

WallFacer: Guiding Transformer Model Training Out of the Long-Context Dark Forest with N-body Problem

Ziming Liu
liuziming@comp.nus.edu.sg
National University of Singapore
The Republic of Singapore

Shaoyu Wang
shaoyu_wang@hust.edu.cn
Huazhong University of Science and
Technology
People's Republic of China

Shenggan Cheng
shenggan@comp.nus.edu.sg
National University of Singapore
The Republic of Singapore

Zhongkai Zhao
zhongkai.zhao@u.nus.edu
National University of Singapore
The Republic of Singapore

Xuanlei Zhao
xuanlei@comp.nus.edu.sg
National University of Singapore
The Republic of Singapore

James Demmel
demmel@berkeley.edu
University of California, Berkeley
The United States of America

Yang You
youy@comp.nus.edu.sg
National University of Singapore
The Republic of Singapore

Abstract

In recent years, Transformer-based Large Language Models (LLMs) have garnered significant attention due to their exceptional performance across a variety of tasks. However, training these models on long sequences presents a substantial challenge in terms of efficiency and scalability. Current methods are constrained either by the number of attention heads, limiting scalability, or by excessive communication overheads. In this paper, we propose an insight that Attention Computation can be considered as a special case of n-body problem with direct interactions. Based on this concept, this paper introduces WallFacer, an efficient long-sequence training system with a novel multi-dimensional ring sequence parallelism, fostering an efficient communication paradigm and extra tuning space for communication arrangement. Through comprehensive experiments under diverse environments and model settings, we demonstrate that WallFacer significantly surpasses state-of-the-art method that supports near-infinite sequence length, achieving performance improvements of up to 77.12%.

Keywords: distributed deep learning, sequence parallelism, n-body simulation, large scale training, high performance computing

1 Introduction

Over the past decade, deep learning has made remarkable strides in diverse fields, including computer vision (CV) and natural language processing (NLP). Transformer-based [50] models, leveraging the groundbreaking attention mechanism introduced in 2018, have secured a dominant position due to their superior feature-capturing capabilities. As the technology has evolved, the ability to efficiently process long

sequences has emerged as a pivotal challenge, capturing the attention of researchers and practitioners alike.



Figure 1. Comparison of the theoretical peer-to-peer communication amount for Ring Attention and different configurations of WallFacer while training a llama-7B model[49] on 64 GPUs with Adam optimizer[22] and a batch size of 4.

This focus on long-sequence training and inference is underscored by its critical importance in a variety of downstream tasks. For instance, in text summarization, the ability to handle extensive sequences is vital, as the content to be summarized can range from lengthy chapters to entire books [6, 24]. Similarly, chat-based applications, such as ChatGPT [1], require the capacity to process extensive dialogue histories to ensure conversational consistency. There are also applications in other fields like video generation[8, 41] and protein structure prediction[10, 21]. In summary, the demand for efficient long-sequence training and inference mechanisms spans a broad spectrum of applications, highlighting a critical area of need within the community.

The long context in the above scenarios has introduced several challenges for model training and inference: 1) **Efficiency and Adaptability**. The challenge of efficiency is predominantly determined by computation and communication. How to maintain high efficiency in various environments and settings has drawn great attention from academia and industry. 2) **Memory**. As the size of the activation during attention is quadratically related to the length of the sequence, memory consumption increases rapidly with the growth of the sequence. 3) **Scalability**. Large Language Models (LLMs) necessitate the use of thousands of GPUs for training, even with datasets of regular lengths. For long sequences, ensuring an acceptable scaling speedup rate is even more critical to reduce time and economic costs.

Traditional parallelisms such as Data Parallelism[17, 28, 46, 53], Tensor Parallelism[46, 51, 52], and Pipeline Parallelism[14, 19, 29, 33] have not been able to address the large memory requirement of extremely long sequences. To break through this obstacle, Sequence Parallelism has been introduced, splitting the input on the sequence length dimension. Mainstream Sequence Parallelism schemes can generally be classified into two categories: those based on all-to-all communication, and those based on ring peer-to-peer communication. Methods like DeepSpeed Ulysses[20], which are based on all-to-all communication, offer efficiency but require the splitting of attention heads. Consequently, these methods are limited in scalability. On the other hand, peer-to-peer communication methods[27, 31], such as Ring Attention[31], do allow for infinite context lengths; however, they necessitate the transmission of complete keys and values across all GPUs, leading to significantly high communication loads. In summary, there remains a deficiency in communication-efficient methods that are capable of supporting infinite context lengths.

In fact, we can draw valuable insights from other fields that have extensive experience in handling long sequences. In the field of n-body simulations, researchers have been dealing with vast numbers of particles, typically far exceeding the number of tokens addressed in current Transformer models. Through careful observation, we have identified that the core component of Transformers, the attention mechanism, closely resembles a special case of the n-body problem. Specifically, n-body problems with direct interactions focus on calculating the resultant force exerted by all other particles in the system on each individual particle. Similarly, in self-attention, the objective is to compute the attention score that each token in the sequence assigns to every other token. This similarity suggests that we can leverage methodologies from n-body problem research—refined over decades—to inform and enhance our approaches to tackling the relatively new challenges of long-context Transformer training.

Inspired by the methodology of n-body communication optimization[13] and tailored to the Transformer architecture, we introduce WallFacer, a near-infinite-context Transformer training system with multi-ring sequence parallelism

that incorporates an additional parallel dimension to the existing ring-style communication, fostering an efficient communication paradigm and extra tuning space for communication arrangement. With very little extra cost of memory, WallFacer parallelism significantly reduces the peer-to-peer communication amount, as is shown in figure 1.

In summary, our paper presents these contributions:

- We conceptualize Attention computation as a novel instance of the traditional n-body problem, providing fresh insights into optimizing and parallelizing Attention computation.
- We introduce a near-infinite-context training system for Transformer models, featuring a groundbreaking multi-ring sequence parallelism scheme. This scheme adds an additional dimension of parallelism and significantly reduces peer-to-peer communication, all while maintaining a minimal memory footprint.
- We offer a straightforward method that allows users to select the most suitable parallelism scheme based on their specific needs, maximizing the utility of the available tuning space within our communication framework.
- We perform experiments on mainstream Transformer models, conducting performance and scaling tests across various computing clusters. Preliminary results indicate that our WallFacer system outperforms Ring Attention by up to 77.12%, showcasing its efficacy and scalability.

2 Background

2.1 Transformer and Long Sequence Training

The key mechanism behind Transformer-based models is Attention[50], which captures the text feature by calculating the attention score between every two single tokens. A standard attention function is given as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where Q, K, and V are the Query, Value, and Key of input x with shape (B, N, H) in the case of self-attention, and d_k represents the head dimension in commonly used multi-head attention. For convenience, we use the symbolic representations in Table 1. We can observe that in the original attention function, we need to store the intermediate value QK^T of size N^2 , which means that the memory capacity requirement and computation amount both grow in quadratic proportion with the sequence length. To solve the memory IO and capacity bottleneck, methods like flash-attention[11] have been proposed. Flash-attention is designed upon the idea of softmax decomposition [23, 36, 42], enabling online softmax computation by storing some extra statistics. Flash-attention is thus capable of segmenting the Q, K, and V matrices into blocks, allowing the entire computation to be performed within a

single kernel for each block sequentially. This strategy effectively circumvents redundant read and write operations to the High Bandwidth Memory (HBM). However, when the sequence length reaches millions, it becomes necessary to distribute the sequence across multiple GPUs. This distribution helps to reduce both the memory and computation demands on any single device. This strategy is also known as Sequence Parallelism. We will explore these methods in further detail in the following section.

Table 1. Meanings of the symbols that are used in this paper

P	The number of GPUs
C	The parallel size of WallFacer (attention parallel size)
H	The hidden dimension size of the Transformer blocks
N	The total number of tokens within the whole sequence
B	The training batch size
W	The communication bandwidth between GPUs
L	The communication latency between GPUs

2.2 Sequence Parallelism

Sequence Parallelism was first conceptualized in a study published in 2021 [30], which introduced the concept as a strategy to distribute computational sequences across multiple processing steps. In this paper, we adopt a broad definition of Sequence Parallelism to include any approach that partitions a sequence during specific computational phases. Presently, Sequence Parallelism can be divided into two main categories: attention-head-sharding-based and peer-to-peer-communication-based. The former involves distributing the attention heads of multi-head attention across multiple GPUs, whereas the latter resembles a distributed version of flash-attention, relying on peer-to-peer communication to transfer keys, values, and intermediate statistics. We will now explore the advantages and disadvantages of these methods in detail.

2.2.1 Attention-head-Sharding-Based. Here we introduce the two representative methods, DeepSpeed Ulysses and Megatron Sequence Parallelism.

DeepSpeed Ulysses. DeepSpeed Ulysses[20] presents a strategy for training with long sequences by leveraging all-to-all collective communication, but is largely limited in its scalability. Functions outside the attention layer are partitioned along the sequence dimension using conventional methods. Within the attention block, Ulysses transitions from sequence parallelism to a method akin to tensor parallelism. It divides the query, key, and value matrices across the attention heads, thereby preserving the original attention computation structure. This is facilitated by two sets of all-to-all communication that switch between sequence splitting and attention head splitting.

The principal merit of Ulysses lies in its simplicity and ease of implementation. Nonetheless, its reliance on the number of attention heads for partitioning activations introduces the following limitations. 1) Ulysses is limited in scalability, as it can only be expanded to as many GPUs as there are attention heads, which in turn restricts the maximum sequence length that can be processed. This limitation becomes particularly problematic when employing techniques like grouped-query attention (GQA) [4] or multi-query attention (MQA) [45], which further reduce the number of available attention heads for Keys and Values. 2) Ulysses encounters load-balancing issues when the number of attention heads is not evenly divisible by the number of GPUs, complicating its deployment across diverse hardware configurations. 3) The need for head-splitting makes Ulysses hard to combine with Tensor Parallelism.

Megatron Sequence Parallelism. Megatron Sequence Parallelism[25] shares its name with an earlier study[30]. Megatron Sequence Parallelism focuses on minimizing memory usage and reducing the necessity for activation recomputation. The training approach is based on Tensor Parallelism (TP). While TP is employed in the Linear and Self-Attention blocks, operations such as the LayerNorm and Dropout functions are not distributed, leading to replications within the tensor parallel group. This requires an additional all-gather operation and a further reduce-scatter operation during the forward propagation to alternate between Tensor Parallelism and Sequence Parallelism. Megatron Sequence Parallelism is not designed for training with exceedingly long sequences, as the Self-Attention and MLP layers continue to rely on Tensor Parallelism and must process the full sequence length. Like DeepSpeed Ulysses, it is also limited by the number of attention heads.

2.2.2 Ring-peer-to-peer-communication-based. The primary method in peer-to-peer-communication-based strategies is Ring Attention[31].

Ring Attention. Introduced in 2023, Ring Attention[31] innovatively partitions the sequence dimension and utilizes a ring-style peer-to-peer (P2P) communication pattern to transfer Keys and Values across all GPUs. Each GPU receives the key and value matrices from the preceding rank, updates the local attention score, and then forwards them to the next rank, as is shown in Figure 2. This method employs an online-softmax and updates attention scores incrementally, allowing the computation of attention scores without retaining the full sequence length. Thus, it potentially supports infinite context, provided sufficient computing resources are available. However, the requirement for P steps of P2P communication renders this approach less efficient in environments with high-latency communication. A more detailed discussion on this topic is presented in the subsequent section.

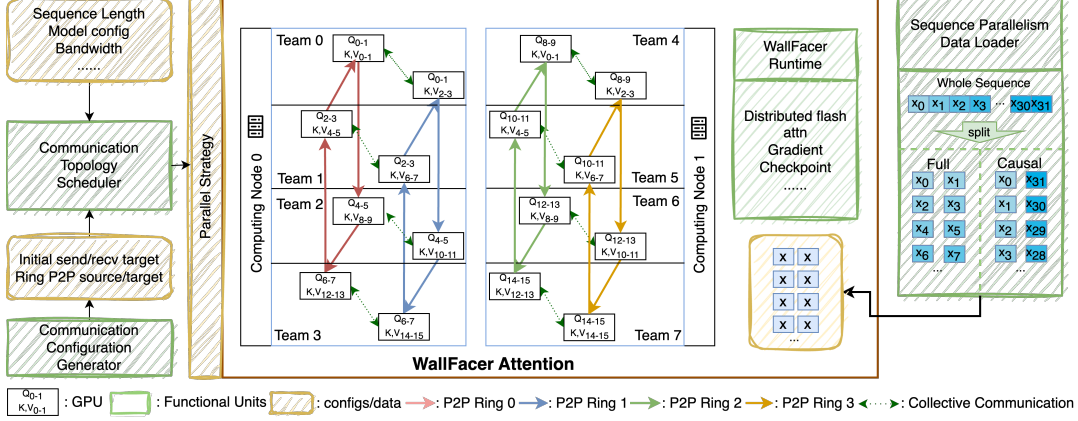


Figure 3. An overview of the WallFacer Training System

evidenced by existing techniques shared between these fields. For instance:

Particle Decomposition: Commonly used to manage memory and computation constraints, this strategy involves distributing particles across multiple processors and integrating the results using the law of force composition. This method is widely implemented in models that employ online softmax, such as Flash-Attention[11].

Hierarchical Methods: Hierarchical structures, such as the Multilevel Fast Multipole Method[34, 48] or tree-based processes like the Barnes-Hut simulation[5], are used to approximate n-body calculations. These methods reduce the complexity from $O(n^2)$ to $O(n \log n)$ or $O(n)$. Such approaches have been effectively applied in various fields, including computer vision, as seen in FASTERVIT[16], etc., and in natural language processing tasks, as demonstrated in [32] and so on.

Cutoff: Interactions in many systems decay with distance, making it practical to set a cutoff distance and ignore interactions beyond this threshold. This principle is similarly utilized in approximate attention algorithms like Longformer[6], which employs a sliding window to calculate an approximate attention score for nearby tokens while excluding those outside the window. These examples validate the feasibility of guiding Transformer system design using n-body simulation techniques. Furthermore, we have identified additional aspects of n-body simulation that could inspire innovative approaches in training long-sequence Transformer models.

In addition to the aforementioned techniques, we have explored a series of **multi-dimensional algorithms** used but not limited in n-body problems[2, 3, 12, 15, 47] that optimize the communication scheme among processors while significantly reducing the total communication load, albeit at the expense of increased memory usage. The core idea involves leveraging an additional parallel dimension to partition the complete communication and computation tasks,

distributing them across various devices. Inspired by these developments, we have conceived the concept of multi-ring sequence parallelism for our training system, which will be discussed in detail in the following section.

4 WallFacer Training System

4.1 Overview

Built upon the inspiration from n-body simulation, the WallFacer training system is comprised of five main components. At the core of the system is WallFacer Attention, which leverages multiple ring-style P2P (peer-to-peer) communication strategies to enhance the efficiency of distributed attention computation. The Dataloader is designed to organize tokens within each sub-sequence according to their mask (causal or full) and distribute them across different GPUs. The Communication Configuration generator plays a critical role in initially assigning Keys and Values to their corresponding ranks. Meanwhile, the Communication Topology Scheduler outlines the placement of parallelism across various computing nodes. Finally, the WallFacer Runtime incorporates additional supporting techniques for the training process, such as gradient checkpointing. We will now explore the details of these components in depth.

4.2 WallFacer Attention

As discussed in the previous section, a major limitation of Ring Attention is the extensive amount of peer-to-peer (P2P) communication required, which becomes problematic in environments with weak connections between computing nodes. Drawing inspiration from the multi-dimensional parallelism used in n-body simulations, we enhance the ring sequence parallelism by introducing an additional dimension. This is achieved by duplicating the Queries, Keys, and Values, thus dividing the communication tasks and distributing them within each team. This segmentation markedly improves computational efficiency and scalability.

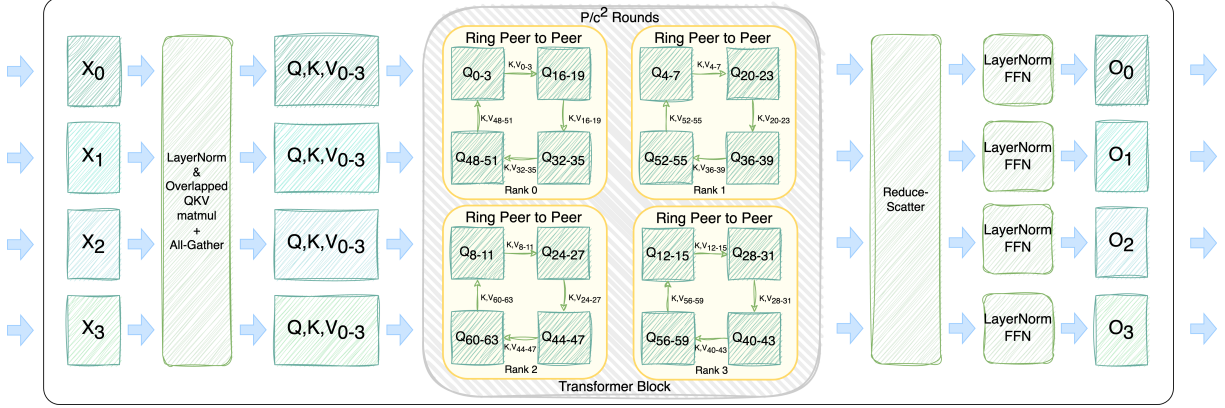


Figure 4. An example of a Transformer block with WallFacer Attention on one team of four devices out of 64 GPUs.

We will now delve into the specifics of the WallFacer Parallelism and provide a theoretical analysis demonstrating its advantages over Ring Attention.

4.2.1 Training Process of WallFacer Parallelism. In the WallFacer system, GPUs are grouped into *Teams* to coordinate computation and communication tasks more effectively, unlike the isolated operations in traditional settings. The Queries, Keys, and Values are gathered within the team, so each member of the team holds the same activation of the whole team before attention. WallFacer introduces an additional parameter, C , which determines the replication factor of the input and, consequently, the number of GPUs within each team. The range of C is from 1 to \sqrt{P} . When C equals one, the algorithm falls back to Ring Attention. When C equals \sqrt{P} , the algorithm becomes a completely collective-communication-based one. When $1 < C < \sqrt{P}$, it becomes a structure with multiple rings looping concurrently. There are three main distinctions between the operational processes of WallFacer and Ring Attention: QKV_matmul & all-gather, attention iteration, and reduce-scatter. Throughout the attention process, asynchronous communication is employed alongside the early launch of communication kernels to maximize the overlap of computation and communication tasks. The complete forward process of a WallFacer Transformer block is outlined in Algorithm 1.

Forward Propagation. In Figure 4, we have an example of one team of four GPUs out of all the 64 GPUs performing WallFacer-style attention. Each training iteration begins with the dataloader splitting the entire input sequence of length N into N/P sub-sequences, which are then loaded onto each GPU. As previously mentioned, the next step involves computing the Queries, Keys, and Values. These are computed separately via matrix multiplication, followed immediately by the launch of the all-gather kernel, which gathers the above QKVs within the team, allowing for the overlap of up to two-thirds of the communication with computation.

Algorithm 1 WallFacer Attention Block (Forward)

Require: Input sequence \mathbf{x} , Linear Function **query**, **key**, and **value**, attention parallelism size c , global rank r , global size gs , team process group pg

- 1: compute the gathered $\mathbf{q}_{team}, \mathbf{k}_{team}, \mathbf{v}_{team} = \text{AllGather_QKVmatmul}(\text{query}, \text{key}, \text{value}, \mathbf{x}, pg)$
- 2: compute the initial rank to send $\mathbf{r}_{send} = \text{get_init_send}(r)$
- 3: compute the initial rank to receive from $\mathbf{r}_{recv} = \text{get_init_recv}(r)$
- 4: launch the asynchronous send and receive request \mathbf{req}_{send} and \mathbf{req}_{recv} , sending $\mathbf{k}_{team}, \mathbf{v}_{team}$ to \mathbf{r}_{send} and receiving $\mathbf{k}_{next}, \mathbf{v}_{next}$ from \mathbf{r}_{recv}
- 5: get the ring P2P target \mathbf{r}_{next} and \mathbf{r}_{last} with $\text{get_P2P_ranks}(r, gs, c)$
- 6: initialize attention score \mathbf{O} , extra statistics \mathbf{lse} to zero.
- 7: **for** $1 \leq i \leq \text{world_size}/c^2$ **do**
- 8: wait for \mathbf{req}_{send} and \mathbf{req}_{recv}
- 9: $\mathbf{k}_{current} = \mathbf{k}_{next}, \mathbf{v}_{current} = \mathbf{v}_{next}$
- 10: launch \mathbf{req}_{send} to send $\mathbf{k}_{current}$ and $\mathbf{v}_{current}$ to \mathbf{r}_{next} , launch \mathbf{req}_{recv} to receive \mathbf{k}_{next} and \mathbf{v}_{next} from \mathbf{r}_{last}
- 11: calculate $\mathbf{lse}, \mathbf{O} = \text{forward_iteration}(\mathbf{lse}, \mathbf{O}, \mathbf{q}_{team}, \mathbf{k}_{current}, \mathbf{v}_{current})$
- 12: **end for**
- 13: compute $\mathbf{O}_{final} = \text{ReduceScatter_combine}(\mathbf{lse}, \mathbf{O}, pg)$
- 14: **return** \mathbf{O}_{final}

Once this phase is complete, each GPU within the team possesses the same Q, K , and V s, each of a length of $\frac{CN}{P}$. To distribute the communication and computation tasks among the team members, we divide the original workload based on four specific ranks assigned to each GPU by the Communication Configuration Generator. These ranks determine each GPU's partners and position within the P2P ring, details of which will be discussed in Section 4.3.

Following the setup, the Keys and Values are dispatched to their designated locations within the cluster to establish the initial sub-ring, setting the stage for the multi-ring iteration phase of WallFacer attention. Given that each sub-sequence is $\frac{CN}{P}$ long and each GPU is tasked with computing the attention score for $\frac{1}{C}$ of the whole sequence, it results in $\frac{N/C}{CN/P} = P/C^2$ rounds of communication. This implies that there are P/C^2 GPUs in one ring.

The iteration process, conducted in a Flash-Attention style, involves storing the log-sum-exp (lse) and intermediate output O , which are updated step by step. Queries are retained locally, while Keys and Values circulate through the ring via P2P communication. After completing the iterations, each team member accumulates the attention scores for the entire team's sub-sequence of Queries with $1/C$ of the Keys and Values from the full sequence.

A simple reduce-scatter operation is then employed to amalgamate the intermediate results and distribute them among the team members. Each GPU ultimately contains the final attention score for its portion of the sequence over the entire sequence. The output of this forward attention block is finalized after a standard LayerNorm and FeedForward layer process.

Backward Propagation. The major distinction between backward and forward propagation, as outlined in section 3, is the inability to calculate Queries independently during the backward phase. Unlike forward propagation, the backward phase requires the complete set of Keys and Values to calculate the gradient for Queries, and vice versa. To manage this, we have structured the gradient calculation into two loops: the Key & Value outer loop and the Query inner loop.

In the outer loop, gradients for Keys and Values are tracked and maintained fixed on the corresponding GPUs within the sub-rings; these gradients do not transfer between GPUs. The inner loop, however, handles the gradients for Queries, which start initialized as zero and are circulated along the sub-rings together with the Queries themselves. During each iteration, the approach mirrors the backward computation method used in Flash Attention, where the updated gradient of the current Query shard is passed to the next GPU in the ring, while the gradients for Keys and Values are retained for subsequent Query shards.

Initial communication ranks for sending and receiving are still configured by the Communication Configuration Generator. Additionally, there is an extra P2P communication required at the end of the process to send the Query and its gradient back to their original location following the loop.

4.2.2 Theoretical Analysis. In this section, we will discuss two major questions: how much communication reduction can WallFacer bring compared with Ring Attention, and does this come with a high extra memory cost? During the analysis, we will employ a case study using the WallFacer system with an attention parallel size of $C = 4$ on a llama-30B

model, which consists of 64 layers. For this model, referred to as model M, the batch size = B is set to 1, the sequence length = N to 65,536, the hidden dimension = H to 6,656, and the number of GPUs = P to 64. Additionally, the computation will utilize bfloat16 precision.

Communication Analysis. Let's analyze the communication overhead within one forward Transformer block on a single GPU. For Ring Attention, the communication is primarily due to the ring P2P loop. Each iteration's communication overhead can be calculated as follows:

$$\frac{2BNH}{PW} + L \quad (4)$$

and as the total number of iterations done is P , the total communication overhead will be:

$$\frac{2BNH}{W} + PL \quad (5)$$

and this overhead can be partially overlapped with the attention computation.

For WallFacer, the communication overhead comes from collective and P2P both. The collective overhead for all-gather and reduce-scatter is:

$$\frac{4BNH(C-1)}{PW} \quad (6)$$

while the P2P communication can be similarly computed as:

$$\frac{P}{C^2} \left(\frac{2CBNH}{PW} + L \right) = \frac{2BNH}{CW} + \frac{P}{C^2} L \quad (7)$$

The advantages of WallFacer over Ring Attention during the ring-P2P phase are evident in three main aspects: 1) **Reduced Communication and Latency:** Ring Attention requires C times more communication than WallFacer, significantly increasing the bandwidth requirement across the entire cluster. For the llama 30B model M, the total communication volume of ring P2P communication and collective communication volume for Ring Attention and WallFacer can be computed as 1.625 GB and 0.152 GB(collective) + 0.406GB (P2P) = 0.558GB. Furthermore, while Ring Attention necessitates P iterations per attention block, WallFacer only requires $\frac{P}{C^2}$, reducing the latency overhead by C^2 . 2) **Localized Communication:** In scenarios like those depicted in Figure 3, WallFacer's ring P2P communication can be confined within the same computing node, where bandwidth is typically much higher than between computing nodes. Conversely, Ring Attention demands inter-node communication during every iteration, which can be less efficient. 3) **Enhanced Overlap of Communication and Computation:** During each iteration, the communication volume of WallFacer is C times higher than that of Ring Attention, while the computational volume during attention is approximately C^2 times greater. This higher computation-to-communication ratio makes it easier for WallFacer to overlap P2P communication with computation, enhancing overall efficiency.

Additionally, the collective communication in WallFacer is minimal due to: 1) Efficient Overlapping: The all-gather

communication overlaps significantly with the QKV matrix multiplication, and the reduce-scatter communication partially coincides with the final attention score update. 2) Scale Considerations: Compared to the P2P communication column, there is a P in the denominator, which is substantial during large-scale training. This implies that collective communication constitutes a very small portion of the total communication volume.

Memory Analysis. In this section, we estimate the theoretical peak memory requirements necessary to store the model weights, activations, and optimizer states. Our implementation utilizes the Adam Optimizer [22], bfloat16 precision, and Zero-2 optimization [43]. Since the WallFacer architecture does not alter the model weights or the optimizer, we can assume that the memory costs associated with the model and the optimizer remain constant across both methods. We name the memory cost for the model and optimizer as M_{m+o} . As for the activation, we refer to the size of one single activation of a sub-sequence on one GPU as

$$A = \frac{B \times N \times H}{P} \quad (8)$$

As we use the checkpointing scheme from [27], a model of Y layers needs to save $Y + 1$ activations as checkpoints. Now we calculate the approximate peak memory after Q, K, and V are already calculated and before the attention computation at the last layer of the whole model. For Ring Attention and WallFacer, the peak memories are:

$$PM_{Ring} = M_{m+o} + (L + 1)A + 3A = M_{m+o} + (L + 4)A \quad (9)$$

$$PM_{Wall} = M_{m+o} + (L + 1)A + 3CA = M_{m+o} + (L + 3C + 1)A \quad (10)$$

, where C is the WallFacer attention dimension. And for the example model M, the peak memory would be $M_{m+o} + 68A$ and $M_{m+o} + 77A$, and the extra memory cost compared with Ring Attention is less than 13.2%, while the P2P communication volume is reduced by about 75%. In a word, the extra memory cost is acceptable as a tradeoff for the communication reduction.

4.3 Communication Configuration Generator

The Communication Topology Scheduler, as introduced in the previous section, plays a crucial role in the initial setup of Queries, Keys, and Values on GPUs and determining each GPU's position within the sub-rings.

Initially, for the setup stage, it is essential to establish the sub-rings by rearranging the activation positions. Specifically, during forward propagation, the Queries do not require rearrangement; however, the Keys and Values must be transmitted to their corresponding positions in the ring prior to commencing the loop. As illustrated in Figure 5, this initialization ensures that each team member holds a different shard of Keys and Values. Moreover, it guarantees that no two teams within the same ring possess identical Keys and Values.

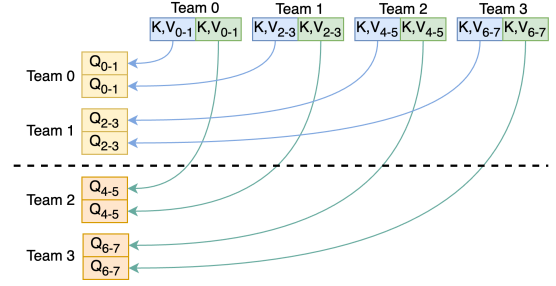


Figure 5. An example of ring initialization process of 8GPUs and 4 sub-rings in the Communication Configuration Generator

Furthermore, Team 0 and Team 1 form what is termed a **team group**, where ring P2P communication occurs exclusively within this group, similarly for Team 2 and Team 3. A team group is defined as a collection of teams that participate in the same sub-rings. The number of teams within a team group can be calculated as $Team\ size = \frac{P}{C}$. We give the details of the determination of initial sending targets in the case of placing teams within same computing nodes in Algorithm 2, and the receiving source can also be calculated similarly.

Algorithm 2 get_init_send()

Require: inter-team rank \mathbf{r}_t , intra-team rank \mathbf{r}_a , inter-team dimension \mathbf{d}_t , intra-team dimension \mathbf{d}_a

- 1: team group size = $\mathbf{d}_t / \mathbf{d}_a$
- 2: target team group rank = \mathbf{r}_a
- 3: target team = target team group rank * team group size + $\mathbf{r}_t // \mathbf{d}_a$
- 4: target device intra-team rank = $\mathbf{r}_t \% \mathbf{d}_a$
- 5: target global rank = target team * \mathbf{d}_a + target device intra-team rank
- 6: return target global rank

After the initialization of activations, we can set up the rings by providing the GPUs their last and next GPU within their rings, as is described in Algorithm 3

4.4 Communication Topology Scheduler

The communication topology scheduler employs a grid-search algorithm to discover the optimal parallelism configuration, tailored to the specifics of the current cluster, model, and input data.

When using Ring Attention, all GPUs form a single ring, which considerably limits the flexibility of adjusting the placement scheme. In contrast, WallFacer expands the tuning space with its unique multi-ring structure in two significant ways. First, the ring configuration in WallFacer is determined by the parallelism dimension $C \in [1, \sqrt{P}]$. Smaller C results

Algorithm 3 get_P2P_config()

Require: inter-team rank r_t , intra-team rank r_a , inter-team dimension d_t , intra-team dimension d_a

- 1: team group size = d_t / d_a
 - 2: self team group rank = $\frac{r_t}{\text{team group size}}$
 - 3: next team in group = $(r_t + 1) \% \text{team group size} + \text{team group size} \times \text{self team group rank}$
 - 4: last team in group = $(r_t - 1) \% \text{team group size} + \text{team group size} \times \text{self team group rank}$
 - 5: next device global rank = $r_a + \text{next team in group} \times d_a$
 - 6: last device global rank = $r_a + \text{last team in group} \times d_a$
 - 7: return next device global rank, last device global rank
-

in less collective communication, a higher total communication volume, and more inter-node communication, while higher C leads to more communication volume during all-gather and reduce-scatter, which may exceed the overhead of the QKV matrix multiplication that we use for overlapping, but also highly reduce the overall communication volume.

Second, the additional dimension provided by WallFacer allows for two strategies in parallelism placement. One strategy is to keep the all-gather and reduce-scatter operations intra-node, which confines these processes within the same computing node. The other strategy is to keep the ring P2P communication intra-node, reducing latency and potentially enhancing performance. The choice between these strategies depends on the extent of overlapping achievable within the ring and the communication volume during collective communications.

Based on the above tuning space, we provide an auto-scheduler, grid-searching through all possible configurations for the one that provides the highest throughput, which can be described as:

$$\text{Config} = \arg \max_{C, \text{placement}} (\text{Profile}(C \in [1, \sqrt{P}], \text{placement} \in [\text{P2P_intra}, \text{Collect_intra}]) \quad (11)$$

The scheduler requires only a few iterations to profile the performance of automatically generated configurations. The time spent on this profiling is negligible compared to the entire training process, as it only needs to be performed once for each new computing cluster used for training.

4.5 Sequence Parallelism Dataloader

To accommodate both full mask and causal mask configurations for self-attention, we have implemented two distinct data loading schemes for load-balancing. For full masks, sequences are straightforwardly divided into equal sub-sequences and distributed among the devices. However, causal masks present a challenge due to the unbalanced computational load across GPUs; sub-sequences at the beginning of a sequence require significantly more computation than those at

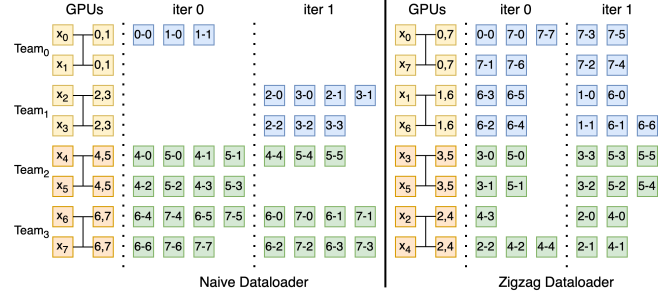


Figure 6. A comparison of using naive and zigzag dataloader for 8 GPUs with attention parallel dimension of 2. The corresponding initialization can be found in Figure 5 with the same configuration. The improvement of efficiency from load-balancing increases with the number of GPUs.

the end. To address this imbalance and achieve load equilibrium among GPUs, we modify the ZigZag scheme introduced by [54], illustrated in Figure 6. The figure illustrates the simplest case of zigzag load-balancing. Notably, the effectiveness of this strategy improves as the number of GPUs increases. This improvement correlates with the expanding difference in computation volume between the first and the last token, which escalates as the sequence length extends. This approach ensures that the total workload on each GPU is balanced, eliminating the need for additional communication mechanisms like those employed in DistFlashAttention[27].

4.6 WallFacer Implementation

WallFacer is written in PyTorch[40] and uses the PyTorch torch.autograd.function and NCCL[39] backend for forward and backward implementation. WallFacer also employs multiple techniques during runtime to improve its overall training efficiency.

Ingestrate Flash Attention. The WallFacer attention mechanism involves multiple iterations that loop over Keys and Values, with each iteration still using traditional self-attention with corresponding Query, Key, and Value (QKV). This approach enables WallFacer to incorporate flash attention effectively, extending its capability by preserving intermediate states across iterations. Additionally, WallFacer enhances the efficiency of the forward process with the help of torch JIT to fuse kernels aside from flash attention.

Overlap communication with computing. In WallFacer attention, P2P communication and self-attention computing are interleaved across iterations, each incurring considerable time. To mitigate this, WallFacer employs a double buffering technique to asynchronously execute communication and computing kernels, effectively overlapping these processes and enhancing GPU utilization.

Save recomputation with checkpoints. WallFacer adopts the checkpointing strategy introduced by DistFlashAttn[27], placing checkpoints at the end of the self-attention phase

rather than the FFN of each transformer layer. This checkpoint placement effectively obviates the need to recompute the self-attention forward process during the backward pass, avoiding redundant attention computation.

5 Evaluation

The computational resources we use in the experiments include a local computing cluster with eight nodes, each equipped with eight Nvidia H100 GPUs with 80GB of high-bandwidth memory, interconnected by NVLink and linked between nodes via InfiniBand with eight Mellanox ConnectX-7 controllers each. Additionally, we utilize Google Cloud Virtual Computing Machines in two setups: one with two machines containing 16 Nvidia A100 GPUs, and another with four machines containing eight GPUs each, both configurations using 40GB of high-bandwidth memory per GPU. All machines are connected within via NVLink and between by Ethernet.

We utilize two different model sizes for the two different GPU types. For H100 with 80GB HBM, we use a GPT model with approximately 7 billion parameters (referred to as 7B in the following text), which has 32 layers, 32 attention heads, and a hidden dimension of 4096. As for the A100 with 40 GB HBM, we use a smaller model with around 3 billion parameters (referred to as 3B), which has 12 layers, 16 attention heads, and a hidden dimension of 4096. During training, both models use bfloat16 precision and a batch size of 1 to accommodate longer input sequences.

In the evaluation section, we aim to answer three major questions:

- How much improvement in throughput can WallFacer bring? Additionally, how adaptable is WallFacer to clusters with both good and poor inter-node connections?
- Is the additional memory cost incurred by WallFacer acceptable considering the throughput improvement it offers?
- How does WallFacer perform in scenarios of weak and strong scaling? Specifically, does it outperform Ring Attention when scaled to handle longer inputs?

5.1 Throughput and Adaptability

Our first experiment aims to assess the performance of WallFacer and Ring Attention across different clusters with varying environments, testing the adaptability of both methods. There are several factors influencing the efficiency of ring-style attention computation:

Theoretical Computation-Communication Volume Ratio: Primarily determined by the sequence length used during training. Attention computation exhibits a computational complexity of $O(N^2 \cdot H)$, whereas P2P communication complexity is $O(N \cdot H)$. Thus, the model configuration does not impact this ratio; only the sequence length does. A larger

N increases the computation-communication ratio, facilitating easier overlap of communication with computation. We evaluate varying sequence lengths to explore performance under different ratios.

Compute Capability and Connectivity of GPUs: The computing overhead, given a specific volume, affects the computation-communication overhead ratio. Higher compute capabilities make overlapping more challenging. We utilize two sets of GPUs in this evaluation: Nvidia A100 40GB and Nvidia H100 80GB, with the latter offering significantly higher theoretical tflops on bf16 computations. Connectivity is considered in two parts: intra-node and inter-node. Our clusters are equipped with NVLink, ensuring robust intra-node communication. For inter-node communication, while InfiniBand offers high bandwidth at a higher commercial cost, some clusters utilize Ethernet for connectivity. Specifically, our H100 nodes leverage InfiniBand with eight adapters per node for superior inter-node bandwidth, whereas the Google Cloud servers use Ethernet. The diversity in node configurations (8-GPU and 16-GPU nodes) allows us to assess adaptability across different topologies.

This evaluation not only highlights the inherent differences between the schemes but also tests their flexibility in various hardware settings.

The results of our evaluation are illustrated in Figure 7. We measure throughput in thousands of tokens per second. To better demonstrate how to select the optimal configuration of WallFacer under each condition, we included two configurations, Wall-2 and Wall-4, in the figure. We omit configurations with lower performance for clarity. As indicated in the figure, in all six settings, at least one configuration of WallFacer achieves higher throughput than Ring Attention, with performance improvements of 62.87%, 42.45%, 35.98%, 19.9%, 77.12%, and 34.62%. This advantage is primarily due to the additional parallel dimension that WallFacer introduces. Unlike Ring Attention, which requires inter-node P2P communication in each iteration, WallFacer’s P2P communication is mostly confined intra-node, except for initial data transfers. This experiment clearly demonstrates WallFacer’s superior performance across various environments.

Another observation is that the optimal configuration for WallFacer may vary depending on the environment, reflecting differences in the computation-communication ratio and the trade-offs between collective and P2P communication. For example, the best C value for A100_16 is 2, while for A100_8, it is 4. This variation can be attributed to the topological differences: the 16-GPU-node cluster benefits less from reduced P2P communication volume, making a smaller C preferable to decrease collective communication volume while maintaining P2P communication at an acceptable level. This finding underscores the importance of the Communication Topology Scheduler we provide, which helps users avoid the complex process of manually analyzing these factors.



Figure 7. Throughput evaluation of Ring Attention and WallFacer on 32 GPUs from three different clusters. We place the performance of WallFacer with both $C=2$ and $C=4$ in the figure. The configurations are marked in the titles of the sub-figures. For instance, A100_8(3B, 63K) represents that the experiment is on machines with 8 Nvidia A100 GPUs in each node, the model used has three billion parameters, and the sequence length is 64k.

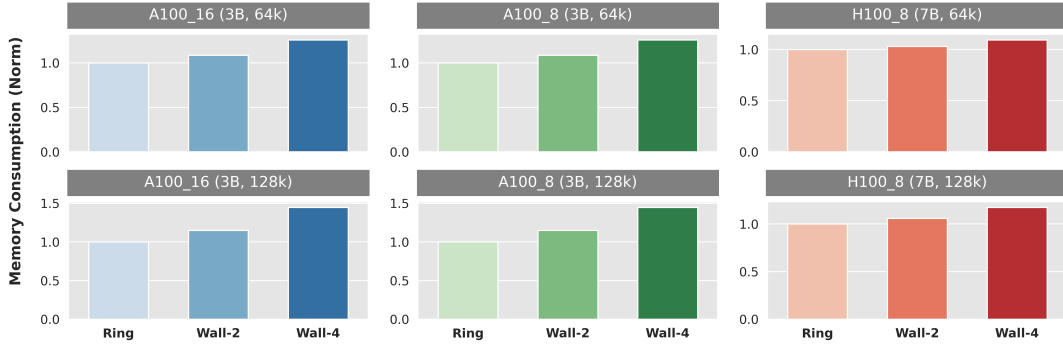


Figure 8. The normalized relative memory cost of different configurations of WallFacer compared with Ring Attention on different clusters.

In summary, WallFacer demonstrates superior efficiency and adaptability in all six cases, thanks to its flexible parallelism scheme.

5.2 Memory Consumption

Memory consumption in our experiments stems from both the model weights and activations, with additional costs for WallFacer arising from the duplication of QKV matrices prior to attention computation. To quantitatively assess the additional memory required by WallFacer, we monitored the maximum memory allocated by PyTorch [40] during our experiments. It is important to note that memory fragmentation in PyTorch can impact memory allocation efficiency. To mitigate this, we limited the sequence lengths in our training to prevent PyTorch from triggering `cuda_free` when the

allocated memory approaches the GPU’s limit, which would otherwise introduce significant overhead.

The results, displayed in Figure 8, reveal that for the configurations yielding the highest throughput, Ring Attention consumes between 7.9% and 30.79% less GPU memory than WallFacer. However, considering the substantial throughput gains provided by WallFacer, the additional memory usage is deemed acceptable. Additionally, evaluations for long sequences typically employ sequence lengths that are powers of two, and we observed that the maximum supported sequence lengths were not significantly impacted by this additional memory usage.

Moreover, in scenarios involving larger models with thirty billion (30B) or seventy billion (70B) parameters, the relative increase in memory consumption due to QKV duplication diminishes. This reduction is explained by equation 10, which

shows that as the model size increases—owing to more parameters and additional Transformer layers—the proportion of memory consumed by model weights and activations grows, whereas the absolute extra memory incurred by QKV multiplication remains constant. This phenomenon is further evidenced by the fact that the extra memory ratio for experiments with the 7B model is significantly smaller than that for the 3B model.

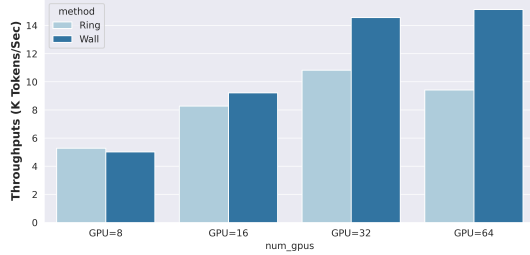


Figure 9. Strong scaling experiments on Nvidia H100 GPUs with fixed sequence length of 128K.

5.3 Strong and Weak Scaling

In the scaling tests we carry out experiments for both strong and weak scaling. Strong Scaling maintains the scale of the problem that we are trying to solve while increasing the computing resource that we use to speed it up. So in this experiment, we fix the sequence length to 128K while increasing the number of GPUs from 8 to 64. As depicted in Figure 9, WallFacer shows gradually more advantage over Ring Attention as we increase the number of GPUs. When we use 64 GPUs, WallFacer shows 60.71% higher throughput than Ring Attention. This can also be explained by the computation-communication ratio. When scaled to more GPUs, the local sequence length on each GPU becomes smaller, and as explained in the previous sections, makes it harder to overlap the P2P communication with attention computation. This is exactly the main advantage of WallFacer, reduction of total P2P communication volume. So WallFacer can give a promising performance when we are trying to speedup the training of a fixed amount of data with more GPUs.

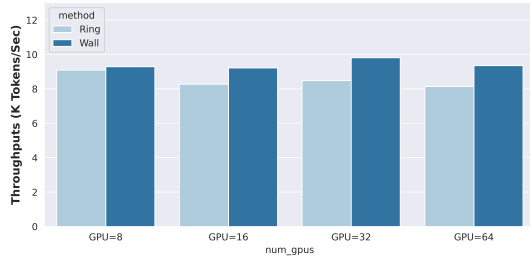


Figure 10. Weak scaling on Nvidia H100 GPUs of sequence length from 64K to 512K

As for weak scaling, we can see from Figure 10 that WallFacer can still show gradually more advantage over Ring Attention, but not as much as that in strong scaling experiments. This is because the computation-communication ratio stays the same in weak scaling. But with more GPUs, Ring Attention will require a higher portion of inter-node communication, while WallFacer can remain most communication local.

In summary, WallFacer shows better scalability in both strong and weak scaling experiments, making it a better choice for large-scale Transformer model training.

6 Related Works

Attention Optimization. Traditional full attention mechanisms necessitate $O(n^2)$ memory for storing the outputs of QK^T , leading to significant computational and memory demands. To address these challenges within the GPU, several approaches have been devised to reduce both memory and computational requirements. Memory-efficient attention[42] introduces a straightforward algorithm that requires only $O(1)$ memory relative to the sequence length, with an extension for self-attention that needs only $O(\log n)$ memory. Flash Attention further minimizes I/O overhead and enhances overall efficiency. Additionally, optimization methods specifically tailored for inference, such as PagedAttention[26], are also being developed to improve the efficiency of attention computations. In this work, we utilize Flash Attention within each iteration to reduce the computation overhead.

Long-Sequence Training Techniques. In recent years, numerous techniques have been developed for long-sequence training. Sequence Parallelism[30] was initially introduced to enhance the efficiency of parallel long-sequence training. Ring Attention[31] improved communication efficiency through memory-efficient methods[42], supporting near-infinite sequence lengths. DeepSpeed Ulysses[20] employs attention head splitting to achieve high efficiency, though it is constrained by the number of heads. Megatron Sequence Parallelism focuses on reducing memory costs during Tensor Parallelism, while DistFlashAttention[27] features a load-balance scheme and a novel gradient checkpoint method. Our work builds on these innovations, introducing a system that supports near-infinite sequence length and large-scale training with an efficient communication scheme.

Techniques for Distributed Model Training. Distributed model training encompasses two primary areas: 1) **Memory Management:** Various techniques aim to conserve GPU memory during distributed training, such as mixed precision training[35] and the ZeRO series[43]. Additionally, Zero-Offload[44] leverages CPU memory offloading. In this work, we implement ZeRO-2 to manage optimizer states and gradients efficiently. 2) **Hybrid Parallelism:** Frameworks like Megatron[37] and Colossal AI[7] integrate multiple forms

of parallelism. There are various existing Parallelism techniques like Pipeline Parallelism[14, 19, 29, 33] and Tensor Parallelism[46], which can be combined with WallFacer Parallelism to facilitate large-scale training. We are also considering the integration of additional frameworks such as [9] to enhance overlapping capabilities in future implementations. In the future, we will examine hybrid strategies that combine WallFacer with other parallelisms.

7 Conclusion

WallFacer represents an advanced near-infinite-context Transformer model training system, featuring a communication-optimized multi-ring sequence parallelism scheme. In this study, we also conceptualize attention computation as a special case of the n-body problem, which introduces new perspectives and potential solutions for future challenges. Through comprehensive adaptability and scaling experiments, we demonstrate that our system not only achieves high efficiency across various training environments but also excels under both strong and weak scaling conditions. In an era increasingly demanding longer contexts for both natural language processing and computer vision, WallFacer is poised to make significant contributions to the industry and inspire innovative research in academia.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [2] R. C. Agarwal, S. M. Balle, F. G. Gustavson, M. Joshi, and P. Palkar. 1995. A three-dimensional approach to parallel matrix multiplication. *IBM Journal of Research and Development* 39, 5 (1995), 575–582. <https://doi.org/10.1147/rd.395.0575>
- [3] Alok Aggarwal, Ashok K Chandra, and Marc Snir. 1990. Communication complexity of PRAMs. *Theoretical Computer Science* 71, 1 (1990), 3–28.
- [4] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. arXiv:2305.13245 [cs.CL]
- [5] J. E. Barnes and P. Hut. 1986. A hierarchical $O(n \log n)$ force calculation algorithm. *Nature* 324 (1986), 446.
- [6] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150 [cs.CL]
- [7] Zhengda Bian, Hongxin Liu, Boxiang Wang, Haichen Huang, Yongbin Li, Chuanrui Wang, Fan Cui, and Yang You. 2021. Colossal-AI: A Unified Deep Learning System For Large-Scale Parallel Training. *arXiv preprint arXiv:2110.14883* (2021).
- [8] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. <https://openai.com/research/video-generation-models-as-world-simulators>
- [9] Chang Chen, Xiuhong Li, Qianchao Zhu, Jiangfei Duan, Peng Sun, Xingcheng Zhang, and Chao Yang. 2024. Centauri: Enabling Efficient Scheduling for Communication-Computation Overlap in Large Model Training via Communication Partitioning. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3* (<conf-loc>, <city>La Jolla</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) (ASPLOS '24). Association for Computing Machinery, New York, NY, USA, 178–191. <https://doi.org/10.1145/3620666.3651379>
- [10] Shenggan Cheng, Xuanlei Zhao, Guangyang Lu, Jiarui Fang, Zhongming Yu, Tian Zheng, Ruidong Wu, Xiwen Zhang, Jian Peng, and Yang You. 2023. FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours. arXiv:2203.00854 [cs.LG]
- [11] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. arXiv:2205.14135 [cs.LG]
- [12] Eliezer Dekel, David Nassimi, and Sartaj Sahni. 1981. Parallel Matrix and Graph Algorithms. *SIAM J. Comput.* 10, 4 (1981), 657–675. <https://doi.org/10.1137/0210049> arXiv:https://doi.org/10.1137/0210049
- [13] Michael B. Driscoll, Evangelos Georganas, Penporn Koanantakool, Edgar Solomonik, and Katherine A. Yelick. 2013. A Communication-Optimal N-Body Algorithm for Direct Interactions. *2013 IEEE 27th International Symposium on Parallel and Distributed Processing* (2013), 1075–1084. <https://api.semanticscholar.org/CorpusID:3941907>
- [14] Shiqing Fan, Yi Rong, Chen Meng, Zongyan Cao, Siyu Wang, Zhen Zheng, Chuan Wu, Guoping Long, Jun Yang, Lixue Xia, Lansong Diao, Xiaoyong Liu, and Wei Lin. 2020. DAPPLE: A Pipelined Data Parallel Approach for Training Large Models. <https://doi.org/10.48550/ARXIV.2007.01045>
- [15] Himanshu Gupta and P. Sadayappan. 1994. Communication efficient matrix multiplication on hypercubes. In *Proceedings of the Sixth Annual ACM Symposium on Parallel Algorithms and Architectures* (Cape May, New Jersey, USA) (SPAA '94). Association for Computing Machinery, New York, NY, USA, 320–329. <https://doi.org/10.1145/181014.181434>
- [16] Ali Hatamizadeh, Greg Heinrich, Hongxu Yin, Andrew Tao, Jose M. Alvarez, Jan Kautz, and Pavlo Molchanov. 2024. FasterViT: Fast Vision Transformers with Hierarchical Attention. arXiv:2306.06189 [cs.CV]
- [17] W Daniel Hillis and Guy L Steele Jr. 1986. Data parallel algorithms. *Commun. ACM* 29, 12 (1986), 1170–1183.
- [18] Roger W. Hockney and James W. Eastwood. 1989. *Computer simulation using particles*. Hilger.
- [19] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, HyukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. 2018. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. <https://doi.org/10.48550/ARXIV.1811.06965>
- [20] Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2023. DeepSpeed Ulysses: System Optimizations for Enabling Training of Extreme Long Sequence Transformer Models. arXiv:2309.14509 [cs.LG]
- [21] John M. Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andy Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David A. Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (2021), 583 – 589. <https://api.semanticscholar.org/CorpusID:235959867>
- [22] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]
- [23] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. arXiv:2001.04451 [cs.LG]
- [24] Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An Empirical Survey on Long Document Summarization: Datasets, Models, and Metrics. *Comput. Surveys* 55 (2022), 1 – 35. <https://api.semanticscholar.org/CorpusID:250118028>
- [25] Vijay Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Reducing Activation Recomputation in Large Transformer Models. arXiv:2205.05198 [cs.LG]
- [26] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. arXiv:2309.06180 [cs.LG]
- [27] Dacheng Li, Rulin Shao, Anze Xie, Eric P. Xing, Xuezhe Ma, Ion Stoica, Joseph E. Gonzalez, and Hao Zhang. 2024. DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training. arXiv:2310.03294 [cs.LG]
- [28] Shigang Li, Tal Ben-Nun, Salvatore Di Girolamo, Dan Alistarh, and Torsten Hoefer. 2020. Taming unbalanced training workloads in deep learning with partial collective operations. In *Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. 45–61.
- [29] Shigang Li and Torsten Hoefer. 2021. Chimera: efficiently training large-scale neural networks with bidirectional pipelines. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–14.
- [30] Shenggui Li, Fuzhao Xue, Yongbin Li, and Yang You. 2021. Sequence Parallelism: Long Sequence Training from System Perspective. In *Annual Meeting of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:246017095>
- [31] Hao Liu, Matei Zaharia, and Pieter Abbeel. 2023. Ring Attention with Blockwise Transformers for Near-Infinite Context. arXiv:2310.01889 [cs.CL]
- [32] Yang Liu and Mirella Lapata. 2019. Hierarchical Transformers for Multi-Document Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David

- Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 5070–5081. <https://doi.org/10.18653/v1/P19-1500>
- [33] Ziming Liu, Shenggan Cheng, Haotian Zhou, and Yang You. 2023. Hanayo: Harnessing Wave-like Pipeline Parallelism for Enhanced Large Model Training Efficiency. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (Denver, CO, USA,) (SC '23). Association for Computing Machinery, New York, NY, USA, Article 56, 13 pages. <https://doi.org/10.1145/3581784.3607073>
- [34] Cai-Cheng Lu and Weng Cho Chew. 1994. A multilevel algorithm for solving a boundary integral equation of wave scattering. *Microwave and Optical Technology Letters* 7, 10 (1994), 466–470. <https://doi.org/10.1002/mop.4650071013> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/mop.4650071013>
- [35] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed Precision Training. arXiv:1710.03740 [cs.AI]
- [36] Maxim Milakov and Natalia Gimelshein. 2018. Online normalizer calculation for softmax. arXiv:1805.02867 [cs.PF]
- [37] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prithvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM. <https://doi.org/10.48550/ARXIV.2104.04473>
- [38] Frank Nielsen and Ke Sun. 2016. Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities. *Entropy* 18, 12 (2016). <https://doi.org/10.3390/e18120442>
- [39] NVIDIA. 2020. NVIDIA Collective Communications Library. <https://developer.nvidia.com/ncccl>
- [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://doi.org/10.48550/ARXIV.1912.01703>
- [41] William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. arXiv:2212.09748 [cs.CV]
- [42] Markus N. Rabe and Charles Staats. 2022. Self-attention Does Not Need $O(n^2)$ Memory. arXiv:2112.05682 [cs.LG]
- [43] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. arXiv:1910.02054 [cs.LG]
- [44] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. ZeRO-Offload: Democratizing Billion-Scale Model Training. arXiv:2101.06840 [cs.DC]
- [45] Noam Shazeer. 2019. Fast Transformer Decoding: One Write-Head is All You Need. arXiv:1911.02150 [cs.NE]
- [46] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053* (2019).
- [47] Edgar Solomonik and James Demmel. 2011. Communication-Optimal Parallel 2.5D Matrix Multiplication and LU Factorization Algorithms. In *Euro-Par 2011 Parallel Processing*, Emmanuel Jeannot, Raymond Namyst, and Jean Roman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 90–109.
- [48] J. M. Song and W. C. Chew. 1995. Multilevel fast-multipole algorithm for solving combined field integral equations of electromagnetic scattering. *Microwave and Optical Technology Letters* 10, 1 (1995), 14–19. <https://doi.org/10.1002/mop.4650100107> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/mop.4650100107>
- [49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]
- [50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]
- [51] Boxiang Wang, Qifan Xu, Zhengda Bian, and Yang You. 2022. Tesseract: Parallelize the Tensor Parallelism Efficiently. In *Proceedings of the 51st International Conference on Parallel Processing*. 1–11.
- [52] Qifan Xu, Shenggui Li, Chaoyu Gong, and Yang You. 2021. An efficient 2d method for training super-large deep learning models. *arXiv preprint arXiv:2104.05343* (2021).
- [53] Yang You, Zhao Zhang, Cho-Jui Hsieh, James Demmel, and Kurt Keutzer. 2018. Imagenet training in minutes. In *Proceedings of the 47th International Conference on Parallel Processing*. 1–10.
- [54] Zilin Zhu et al. 2024. Ring Flash Attention. [github:https://github.com/zhuilin/ring-flash-attention](https://github.com/zhuilin/ring-flash-attention)