

# LEGO SCALE: ONE-STOP PYTORCH NATIVE SOLUTION FOR PRODUCTION READY LLM PRE-TRAINING

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

The development of large language models (LLMs) has been instrumental in advancing state-of-the-art natural language processing applications. Training LLMs with billions of parameters and trillions of tokens require sophisticated distributed systems that enable composing and comparing several state-of-the-art techniques in order to efficiently scale across thousands of accelerators. However, existing solutions are complex, scattered across multiple libraries/repositories, lack interoperability, and are cumbersome to maintain. Thus, curating and empirically comparing training recipes require non-trivial engineering effort.

This paper introduces LEGOSCALE, an open-source, PyTorch-native distributed training system that unifies and advances state-of-the-art techniques, streamlining integration and reducing engineering overhead. LEGOSCALE enables seamless application of 3D parallelism in a modular and composable manner, while featuring elastic scaling to adapt to changing computational requirements. The system provides comprehensive logging, efficient checkpointing, and debugging tools, ensuring production-ready training. Moreover, LEGOSCALE incorporates innovative hardware-software co-designed solutions, leveraging cutting-edge features like Float8 training and SymmetricMemory to maximize hardware utilization. As a flexible experimental test bed, LEGOSCALE facilitates the curation and comparison of custom recipes for diverse training contexts. By leveraging LEGOSCALE, we developed optimized training recipes for the Llama 3.1 family and provide actionable guidance on selecting and combining distributed training techniques to maximize training efficiency, based on our hands-on experiences.

We thoroughly assess LEGOSCALE on the Llama 3.1 family of LLMs, spanning 8 billion to 405 billion parameters, and showcase its exceptional performance, modular composability, and elastic scalability. By stacking training optimizations, we demonstrate accelerations ranging from 65.08% on Llama 3.1 8B at 128 GPU scale (1D), 12.59% on Llama 3.1 70B at 256 GPU scale (2D), to 30% on Llama 3.1 405B at 512 GPU scale (3D) on NVIDIA H100 GPUs over optimized baselines.

## 1 INTRODUCTION

**LLMs are at the forefront of NLP advancement.** Large Language Models (LLMs) (Devlin, 2018; Liu et al., 2019; Radford et al., 2019; Chowdhery et al., 2023; Anil et al., 2023; Achiam et al., 2023; Dubey et al., 2024; Jiang et al., 2024; Abdin et al., 2024) have been driving force behind the advancement of natural language processing (NLP) applications spanning language translation, content/code generation, conversational AI, text data analysis, creative writing and art, education and research etc.

**LLMs require billions of parameters and training over trillion tokens to achieve state-of-the-art performance.** Achieving state-of-the-art LLM performance requires massive scale, exemplified by top-performing models like Llama 3.1 (405B parameters, 15T tokens, 30.84M GPU hours, 16K H100 GPUs) (Dubey et al., 2024) and Google’s PaLM (540B parameters, 0.8T tokens, 9.4M TPU hours, 6144 TPUv4 chips) (Chowdhery et al., 2023). These models demonstrate exceptional natural language understanding and generation capabilities, but necessitate substantial computational resources, memory, and time to train, highlighting the significant investment required to advance natural language processing.

**LLM training challenges are being tackled from all sides.** Training large language models (LLMs) at scale is a daunting task that requires a delicate balance of parallelism, computation, and communication, all while navigating intricate memory and computation tradeoffs. The massive resources required for training make it prone to GPU failures, underscoring the need for efficient recovery mechanisms and checkpointing strategies to minimize downtime (Eisenman et al., 2022; Wang et al., 2023; Gupta et al., 2024; Maurya et al., 2024; Wan et al., 2024). To optimize resource utilization and achieve elastic scalability, it is crucial to combine multiple parallelism techniques, including Data Parallel (Li et al., 2020; Rajbhandari et al., 2020; Zhang et al., 2022; Zhao et al., 2023), Tensor Parallel (Narayanan et al., 2021; Wang et al., 2022; Korthikanti et al., 2023), Context Parallel (Liu et al., 2023; Liu & Abbeel, 2024; NVIDIA, 2023; Fang & Zhao, 2024), and Pipeline Parallel (Huang et al., 2019; Narayanan et al., 2019; 2021; Tang et al., 2024). By stacking these parallelisms with memory and computation optimization techniques, such as activation recomputation (Chen et al., 2016; Korthikanti et al., 2023; He & Yu, 2023), mixed precision training (Micikevicius et al., 2018; 2022), and deep learning compilers (Bradbury et al., 2018; Yu et al., 2023; Li et al., 2024; Ansel et al., 2024), it is possible to maximize hardware utilization.

**Limitations of existing systems incorporating state-of-the-art techniques.** While state-of-the-art distributed training techniques have significantly advanced the field, existing systems that incorporate them still fall short in addressing critical challenges that hinder their usability, adoption and effectiveness for researchers and industry practitioners.

1. **Non-composable:** Existing systems struggle to combine and stack various parallelism techniques, limiting the exploration of multi-dimensional parallelism. Further integrating them with memory and computation optimizations is challenging, hindering training efficiency.
2. **Inflexible and monolithic architecture:** Current systems are not modular or extensible, making it difficult to integrate and compare new techniques, optimizations, and hardware, and limiting adaptability to evolving machine learning landscapes.
3. **Inefficient hardware utilization:** Current systems fail to fully leverage advanced hardware features, leading to sub-optimal GPU efficiency, and lack customizable activation checkpointing strategies to navigate memory-computation trade-offs.
4. **Insufficient support for production-grade training:** Existing systems lack scalable and efficient distributed checkpointing, making failure recovery and model saving cumbersome, and often do not provide adequate debugging tools and logging metrics, leading to difficulties in identifying and fixing issues, particularly for those without extensive expertise.
5. **Existing systems fall short in harnessing the full potential of frameworks like PyTorch,** missing out on bug fixes, optimized kernels, new features, and compiler support. They also rely on external dependencies that often lack thorough testing and can become outdated or incompatible due to inadequate maintenance.

**Root cause: Lack of an expressive tensor abstraction.** The root cause of non-composability and inflexibility of a distributed system stems from the lack of using an expressive tensor and device abstraction as a central component, upon which all of the distributed parallelisms, checkpointing, and efficiency optimizations can be built.

**Design Principle: Unified distributed tensor and device abstractions as building blocks.** A unified device abstraction represents the distributed system as a multi-dimensional array, where each dimension corresponds to a parallelism technique, managing communication between devices and handling collective process groups. A complementary tensor abstraction enables tensors to be sharded across this array, maintaining sharding specifications and supporting automatic sharding propagation. Together, these abstractions enable seamless composition of parallelism techniques, ensure correct semantics, and facilitate dispatching of collectives for distributed operations

We address the technical challenge of a unified tensor abstraction by employing PyTorch’s Distributed Tensor (DTensor) and DeviceMesh (PyTorch Community, 2023a) as the foundational components for LEGOSCALE. Through our work with DTensor and DeviceMesh, we identified key limitations and addressed them. By using and extending DTensor, we develop LEGOSCALE, a production-ready system that enables composability, modularity, flexibility, and extensibility in distributed training. LEGOSCALE facilitates the composition of 3D parallelism, training optimizations, scalable distributed checkpointing, and harnesses the full benefits of the PyTorch ecosystem.

To develop and evaluate the capabilities of LEGOSCALE, we undertook several key steps, which represent the core contributions of this work, and are summarized as follows:

1. We advance DTensor by extending its sharding to support n-D parallelism, adding compatibility with `torch.compile` for compiler optimizations, and enabling efficient checkpointing of n-D models via state dict support. We also resolve critical bugs to bolster DTensor’s production readiness
2. We demonstrate how to compose and stack various parallelism techniques, facilitating the exploration of multi-dimensional parallelism in large language model training (§2.1).
3. We enable novel hardware-software co-designed solutions exploiting advanced hardware features to increase GPU efficiency, offer customizable activation checkpointing strategies for navigating memory-computation trade-offs, and utilize `torch.compile` to further optimize memory, computation, and communication (§2.2).
4. We offer production grade training by incorporating scalable and efficient distributed checkpointing to facilitate fast failure recovery, integrating debugging tools like Flight Recorder to debug crashed/stuck jobs, and providing extensive logging metrics (§2.3).
5. We extensively evaluate LEGOSCALE on Llama 3.1 family of models, stacking 1D to 3D parallelisms (respectively), at the scale from 8 to 512 GPUs to demonstrate elastic scalability while ensuring efficiency, convergence, and accuracy. In summary, we demonstrate training accelerations ranging from 65.08% on Llama 3.1 8B at 128 GPU scale (1D), 12.59% on Llama3.1 70B at 256 GPU scale (2D), to 30% on Llama3.1 405B at 512 GPU scale (3D) on latest NVIDIA H100 GPUs over optimized baselines (§3.2).
6. We provide systematic training recipes and guidelines that empower users to navigate the complexities of distributed training, helping them optimize training efficiency for a range of model sizes and cluster configurations (§3.3).
7. We show how our modular and extensible architecture allows for seamless integration and comparison of new techniques, optimizations, and hardware, ensuring adaptability to evolving machine learning landscapes (§4).

By providing an accessible and extensible platform, LEGOSCALE democratizes large language model (LLM) pre-training, empowering a wider range of researchers and developers to tap into the potential of LLMs and accelerate innovation in the field.

## 2 ELASTICITY THROUGH COMPOSABILITY

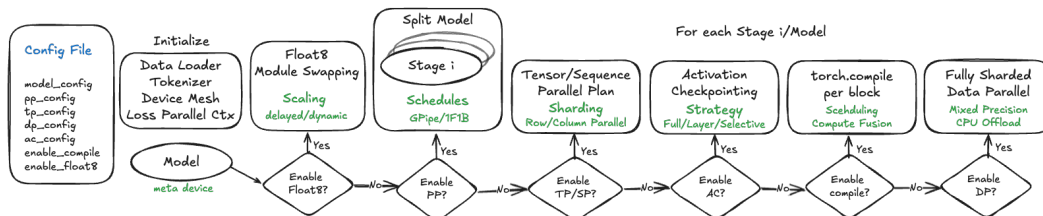


Figure 1: Composable and Modular LEGOSCALE initialization workflow.

LEGOSCALE incorporates various parallelism techniques in a modular manner to enable easy, user-selectable combinations of multi-dimensional parallelisms. This composability enables the tackling of difficult scaling challenges by enhancing the ease of frontier exploration for optimizing training efficiencies at scale.

The codebase of LEGOSCALE is organized purposefully to enable composability and extensibility. We intentionally keep three main components separate and as orthogonal as possible: (1) the model definition, which is parallelism-agnostic and designed for readability, (2) parallelism helpers, which apply Data Parallel, Tensor Parallel, and Pipeline Parallel to a particular model, and (3) a generalized training loop. All these components are configurable via TOML files with command-line overrides, and it is easy to add new models and parallelism techniques on top of the existing codebase.

## 2.1 COMPOSABLE N-D PARALLELISM TRAINING

In this section, we will walk through the entire regime of scaling model training on large clusters, including meta device initialization and the core composable multi-dimensional parallelisms, to showcase how these techniques can be composed to train LLMs efficiently at increasing scale in LEGOSCALE. The corresponding actual code snippets in LEGOSCALE can be found in Appendix A.

### 2.1.1 LARGE-SCALE MODEL INITIALIZATION USING META DEVICE

Given the exponential increase in model sizes for LLMs, the first scaling issue appears even before the actual training starts. This is the need to instantiate a large model for sharding across the cluster, yet without overflowing CPU or GPU memory.

To tackle this, we enabled meta device initialization for models in LEGOSCALE, where the model is first initialized on a “meta” device type. The meta device tensor only holds the metadata information, not the actual data, making initialization ultra-fast. After that, we perform model sharding and transforming the model parameters into Distributed Tensors (DTensors) where each parameter holds a local shard that lives on the meta device. Finally, we perform parameter initialization based on the user-defined initialization functions. We leverage Distributed Tensor to properly sync Random Number Generator (RNG) seeds, and initialize the parameters according to their sharding layouts. This ensures the parameters start with the same values as if the whole model were initialized on one device before sharding, and thus facilitating convergence comparisons between different parallelism configurations.

### 2.1.2 FULLY SHARDED DATA PARALLEL

The original Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023) is an effective implementation of ZeRO that offers large model training capability in PyTorch. However, the original implementation (FSDP1) in PyTorch suffers from various limitations due to its FlatParameter implementation (see details in Appendix B.1).

Given these limitations, LEGOSCALE integrates a new version of Fully Sharded Data Parallel (FSDP2), which uses the per-parameter Distributed Tensor sharding representation and thus provides better composability with model parallelism techniques and other features that require the manipulation of individual parameters,

LEGOSCALE integrates and leverages FSDP2 as its default 1D parallelism, benefiting from the improved memory management (often 7 percent lower per GPU memory requirement vs FSDP1) and the slight performance gains (average of 1.5 percent gain vs FSDP1). More details on FSDP2 and usage example are shown in Appendix B.1. LEGOSCALE makes it simple to run with FSDP2 by embedding appropriate defaults, including auto-sharding with your world size automatically.

For scaling to even larger world sizes, LEGOSCALE also integrates Hybrid Sharded Data Parallel (HSDP) which extends FSDP2 by creating sharding groups. Details are shown in Appendix B.2

### 2.1.3 TENSOR PARALLEL

Tensor Parallel (TP) (Narayanan et al., 2021), together with Sequence Parallel (SP) (Korthikanti et al., 2023), is a key model parallelism technique to enable large model training at scale.

TP is implemented in LEGOSCALE using the PyTorch’s `RowwiseParallel` and `ColwiseParallel` APIs, where the model parameters are partitioned to DTensors and perform sharded computation with it. By leveraging DTensor, the TP implementation does not need to touch the model code, which allows faster enablement on different models and provides better composability with other features mentioned in this paper.

**Tensor and Sequence Parallel (TP/SP)** While TP partitions the most computationally demanding aspects, Sequence Parallel (SP) perform sharded computation for the normalization or dropout layers on the sequence dimension, which otherwise generate large replicated activation tensors, and thus can be challenging to memory constraints per GPU. See Appendix B.3 for more details, illustrations, and usage for both TP and FSDP + TP.

216 Because of the synergistic relationship between TP and SP, LEGOSCALE natively bundles these two  
 217 together and they are jointly controlled by the TP degree setting.  
 218

219 **Loss Parallel** When the loss function is computed, the model outputs are usually very large. Since  
 220 the model outputs from TP/SP are sharded on the (often huge) vocabulary dimension, naively com-  
 221 puting the cross-entropy loss requires gathering all the shards along the TP dimension to make the  
 222 outputs be replicated, which incurs large memory usage.  
 223

224 With Loss Parallel, the cross entropy loss can be computed efficiently, without gathering all the  
 225 model output shards to every single GPU. This not only significantly reduces the memory consump-  
 226 tion, but also improves training speed by reducing communication overhead and doing sharded  
 227 computation in parallel. Given these improvements, LEGOSCALE implements loss parallel by default.  
 228

#### 230 2.1.4 PIPELINE PARALLEL

231 For pre-training at the largest scales, LEGOSCALE offers Pipeline Parallelism, which becomes es-  
 232 sential due to having the lightest amount of communication overhead and leveraging P2P commu-  
 233 nications.  
 234

235 Pipeline Parallel (PP) views the model as a sequence of operations, chunking the operations (and  
 236 the parameters used by them) into  $S$  stages which run on separate groups of devices. In the typi-  
 237 cal case, one stage represents a single model layer or a group of  $N$  adjacent model layers, but in  
 238 theory it could be even be a partial layer. For the forward pass, a stage receives input activations  
 239 (except stage 0), performs local computation, and sends output activations (except stage  $S - 1$ ). The  
 240 last stage performs a loss computation, and begins the backward pass, sending gradients in the re-  
 241 verse order through the pipeline. To improve efficiency, the input batch is broken into microbatches  
 242 and a pipeline schedule overlaps computation on one microbatch with communication for others.  
 243 LEGOSCALE enables a variety of pipeline schedules, with their schedules previously described in  
 244 other works (Narayanan et al., 2019; Huang et al., 2019; Narayanan et al., 2021; Tang et al., 2024).

245 The training loop must also account for creation of pipeline stages, and executing a pipeline sched-  
 246 ule rather than invoking `model.forward()` directly. Because the schedule computes loss per  
 247 microbatch, the loss computation and any logging code must be updated for PP. In LEGOSCALE, we  
 248 propose to define a shared `loss_fn` to be used by both pipeline and non-pipeline code paths, and  
 249 thus minimise divergence in the training loop.

250 Interactions with data parallelism, such as ensuring that data-parallel reduction happens only after  
 251 the last microbatch in the schedule, and scheduling of shard and unshard operations when using  
 252 zero-3, are also handled transparently inside the pipeline schedule executor, simplifying the trainer  
 253 implementation in LEGOSCALE. For it’s usage in LEGOSCALE, please see Appendix B.4.  
 254

## 255 2.2 OPTIMIZING TRAINING EFFICIENCIES

### 257 2.2.1 NAVIGATING COMPUTE-MEMORY TRADE-OFFS USING ACTIVATION CHECKPOINTING

258 Activation checkpointing (AC) (Chen et al., 2016) and selective activation checkpointing (SAC) (Ko-  
 259 rthikanti et al., 2023) are standard training techniques to reduce peak GPU memory usage, by trading  
 260 activation recomputation during the backward pass for memory savings. It is often needed even after  
 261 applying multi-dimensional parallelisms.  
 262

263 LEGOSCALE offers flexible AC and SAC options utilizing `torch.utils.checkpoint`, applied  
 264 at the `TransformerBlock` level. The AC strategies include “full” AC, op-level SAC, and layer-  
 265 level SAC.

266 Within a `TransformerBlock`, full AC works by recomputing all activation tensors needed during  
 267 the backward pass, whereas op-level SAC saves the results from computation-intensive PyTorch  
 268 operations and only recomputes others. Layer-level SAC works in similar fashion as full AC, but  
 269 the wrapping is applied to every  $x$  `TransformerBlock` (where  $x$  is specified by the user) to  
 implement configurable trade-offs between memory and recompute. (Details are in Appendix B.5.)

### 2.2.2 REGIONAL COMPILATION TO EXPLOIT `TORCH.COMPILE` OPTIMIZATIONS

`torch.compile` was released in PyTorch 2 (Ansel et al., 2024) with TorchDynamo as the frontend to extract PyTorch operations into an FX graph, and TorchInductor as the backend to compile the FX graph into fused Triton code to improve the performance.

In LEGOSCALE, we use regional compilation, which applies `torch.compile` to each individual `TransformerBlock` in the Transformer model. This has two main benefits: (1) we get a full graph (without graph breaks) for each region, compatible with FSDP2 and TP (and more generally `torch.Tensor` subclasses such as `DTensor`) and other PyTorch distributed training techniques; (2) since the Llama model stacks identical `TransformerBlock` layers one after another, `torch.compile` can identify the same structure being repeatedly compiled and only compile once, thus greatly reducing compilation time.

`torch.compile` brings efficiency in both throughput and memory (see Section 3.2) via computation fusions and computation-communication reordering, in a model-agnostic way with a simple user interface. Below we further elaborate how `torch.compile` composability helps LEGOSCALE unlock hardware-optimized performance gain with simple user interface, with the integration of advanced features such as Asynchronous TP and Float8.

### 2.2.3 ASYNCHRONOUS TENSOR PARALLEL TO MAXIMALLY OVERLAP COMMUNICATION

By default TP incurs blocking communications before/after the sharded computations, causing computation resources to not be effectively utilized. Asynchronous TP (AsyncTP) (Wang et al., 2022) achieves computation-communication overlap by fractionalizing the TP matrix multiplications within the attention and feed-forward modules into smaller chunks, and overlapping communication collectives in between each section. The overlap is achieved by a micro-pipelining optimization, where results are being communicated at the same time that the other chunks of the matmul are being computed.

PyTorch AsyncTP is based on a `SymmetricMemory` abstraction, which creates intra-node buffers to write faster communication collectives. This is done by allocating a shared memory buffer on each GPU in order to provide direct P2P access.

With LEGOSCALE’s integration of `torch.compile`, AsyncTP can be easily configured in LEGOSCALE to achieve meaningful end-to-end speedups (see Section 3.2 for details) on newer hardware (H100 or newer GPUs with NVSwitch within a node). Usage details are in Appendix B.6

### 2.2.4 BOOSTING THROUGHPUT WITH MIXED PRECISION TRAINING AND FLOAT8 SUPPORT

Mixed precision training (Micikevicius et al., 2018) provides both memory and computational savings while ensuring training stability. FSDP2 has built-in support for mixed precision training with basic `torch.dtype`. This covers the popular usage of performing FSDP all-gather and computation in a low precision (e.g. `torch.bfloat16`), and perform lossless FSDP reduce-scatter (gradient) in high precision (e.g. `torch.float32`) for better numerical results. See Appendix B.7 for usage details.

LEGOSCALE also supports more advanced mixed precision training with Float8 (a derived data type) on newer hardware like H100, with substantial performance gains (reported in Section 3.2). The Float8 feature from `torchao.float8` supports multiple per-tensor scaling strategies, including dynamic, delayed, and static (see Micikevicius et al. (2022); PyTorch Community (2023b), Section 4.3 for details), while being composable with other key PyTorch-native systems such as autograd, `torch.compile`, FSDP2 and TP (with Float8 all-gather capability).

## 2.3 PRODUCTION READY TRAINING

To enable production-grade training, LEGOSCALE offers seamless integration with key features out of the box. These include (1) efficient checkpointing using PyTorch Distributed Checkpointing (DCP), and (2) debugging stuck or crashed jobs through integration with Flight Recorder.

### 2.3.1 SCALABLE AND EFFICIENT DISTRIBUTED CHECKPOINTING

Checkpoints are crucial in training large language models for two reasons: they facilitate model reuse in applications like inference and evaluation, and they provide a recovery mechanism in case of failures. An optimal checkpointing workflow should ensure ease of reuse across different parallelisms and maintain high performance without slowing down training. There are two typical checkpointing methods. The first aggregates the state (model parameters and optimizer states) into an unsharded version that is parallelism-agnostic, facilitating easy reuse but requiring expensive communication. The second method has each trainer save its local sharded state, which speeds up the process but complicates reuse due to embedded parallelism information.

DCP addresses these challenges using DTensor, which encapsulates both global and local tensor information independently of parallelism. DCP converts this information into an internal format for storage. During loading, DCP matches the stored shards with the current DTensor-based model parameters and optimizer states, fetching the necessary shard from storage. LEGOSCALE, which utilizes all native PyTorch parallelisms, effectively uses DCP to balance efficiency and usability. Furthermore, DCP enhances efficiency through asynchronous checkpointing by processing storage persistence in a separate thread, allowing this operation to overlap with subsequent training iterations. LEGOSCALE utilizes DCP’s asynchronous checkpointing to reduce the checkpointing overhead by 5-15x compared to synchronous distributed checkpointing for the Llama 3.1 8B model.

### 2.3.2 FLIGHT RECORDER TO DEBUG JOB CRASHES

A common failure mode when developing parallelism code, or when running at large scale, is to observe a NCCL collective timeout and then the need to figure out the root cause. Since communication kernels are typically asynchronous from the perspective of the CPU, by the time something times out, it can be very hard to pinpoint which operation failed and why. PyTorch provides a Flight Recorder for NCCL collectives to help resolve this dilemma. It records the start and end time (on GPU) as well as the enqueue time (on CPU) for every collective or p2p operation. Additionally, it logs metadata such as which process group was used, who the source rank is (and destination, for p2p), tensor sizes, and stack traces.

We find the data contained in the Flight Recorder helps debug collective hangs and p2p hangs caused by bugs in parallelism code. For PP, there may be schedule bugs that lead to hangs, due to a missing or improperly ordered send or recv operation. Analysis based on the Flight Recorder data can pinpoint the latest send or recv that has completed on the GPU. For FSDP or TP, it is possible to determine whether one or more ranks has not called into a collective, perhaps due to a bug in PP scheduling or faulty logic in TP.

## 3 EXPERIMENTATION

In this section, we demonstrate the effectiveness of elastic distributed training using LEGOSCALE, via experiments on Llama 3.1 8B, 70B, and 405B, from 1D parallelism to 3D parallelism (respectively), at the scale from 8 GPUs to 512 GPUs. We also share the knowledge and experience gained through LEGOSCALE experimentation. A walkthrough of the codebase on how we apply (up to) 3D parallelism can be found in Appendix A.

### 3.1 EXPERIMENTAL SETUP

The experiments are conducted on NVIDIA H100 GPUs<sup>1</sup> with 95 GiB memory, where each host is equipped with 8 GPUs and NVSwitch. Two hosts form a rack connected to a TOR switch. A backend RDMA network connects the TOR switches. In LEGOSCALE we integrate a checkpointable data loader and provide built-in support for the C4 dataset (en variant), a colossal, cleaned version of Common Crawl’s web crawl corpus (Raffel et al., 2020). We use the same dataset for all experiments in this section. For the tokenizer, we use the official one (tiktoken) released together with Llama 3.1.

<sup>1</sup>The H100 GPUs used for the experiments are non-standard. They have HBM2e and are limited to a lower TDP. The actual peak TFLOPs should be between SXM and NVL, and we don’t know the exact value.

## 3.2 PERFORMANCE

To showcase the elasticity and scalability of LEGOSCALE, we experiment on a wide range of GPU scales (from 8 to 512), as the underlying model size increases (8B, 70B, and 405B) with a varying number of parallelism dimensions (1D, 2D, and 3D, respectively). To demonstrate the effectiveness of the optimization techniques introduced in Section 2.2, we show how training throughput improves when adding each individual technique on appropriate baselines. In particular, when training on a higher dimensional parallelism with new features, the baseline is always updated to include all previous techniques.

We note that, throughout our experimentation, memory readings are stable across the whole training process<sup>2</sup>, whereas throughput numbers (token per second, per GPU) are calculated and logged every 10 iterations, and always read at the (arbitrarily determined) 90th iteration. We do not report Model FLOPS Utilization (MFU) (Chowdhery et al., 2023) because when Float8 is enabled in LEGOSCALE, both BFLOAT16 Tensor Core and FP8 Tensor Core are involved in model training, but they have different peak FLOPS and the definition of MFU under such scenario is not well-defined. We note that the 1D Llama 3.1 8B model training on 8 or 128 H100 GPUs without Float8 achieves 33% to 42% MFU.

Table 1: 1D parallelism (FSDP) on Llama 3.1 8B model, 8 GPUs. Mixed precision training. Selective activation checkpointing. Local batch size 2, global batch size 16.

Techniques	Throughput (Tok/Sec)	Comparison	Memory (GiB)
FSDP	6,258	100%	81.9
+ torch.compile	6,674	+ 6.64%	77.0
+ torch.compile + Float8	9,409	+ 50.35%	76.8

Table 2: 1D parallelism (FSDP) on Llama 3.1 8B model, 128 GPUs. Mixed precision training. Selective activation checkpointing. Local batch size 2, global batch size 256.

Techniques	Throughput (Tok/Sec)	Comparison	Memory (GiB)
FSDP	5,645	100%	67.0
+ torch.compile	7,168	+ 26.97%	62.0
+ torch.compile + Float8	9,319	+ 65.08%	61.8

Table 3: 2D parallelism (FSDP + TP) + torch.compile + Float8 on Llama 3.1 70B model, 256 GPUs. Mixed precision training. Full activation checkpointing. FSDP degree 32, TP degree 8. Local batch size 16, global batch size 512.

Techniques	Throughput (Tok/Sec)	Comparison	Memory (GiB)
2D	897	100%	70.3
+ AsyncTP	1,010	+ 12.59%	67.7

Table 4: 3D parallelism (FSDP + TP + PP) + torch.compile + Float8 + AsyncTP on Llama 3.1 405B model, 512 GPUs. Mixed precision training. Full activation checkpointing. FSDP degree 4, TP degree 8, PP degree 16. Local batch size 32, global batch size 128.

Schedule	Throughput (Tok/Sec)	Comparison	Memory (GiB)
1F1B	100	100%	78.0
Interleaved 1F1B	130	+ 30.00%	80.3

<sup>2</sup>Different PP ranks can have different peak memory usages. We take the maximum across all GPUs.



### 3.3 SCALING WITH LEGOSCALE 3D PARALLELISM

The LLM scaling law imposes challenges due to increasingly larger model size and huge amount of data, which requires applying parallelism strategies on massive number of GPUs. LEGOSCALE provides the ability to compose different parallelisms to efficiently scale model training to thousands of GPUs. This section discusses the observations and motivations to apply LEGOSCALE 3D parallelism when training LLMs at large scale. Please note that there could be many 3D parallelism combinations, but we choose to only discuss one combination in this paper, which could be summarized as the following diagram:

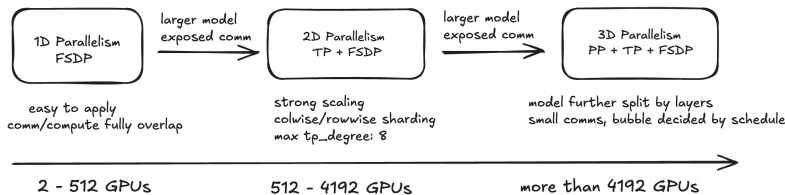


Figure 2: Scaling with 3D Parallelism

#### 3.3.1 SCALING WITH FSDP

FSDP (ZeRO) is a generalized technique which can be applied to any model architecture, making it a good choice for the first or the only degree of parallelism. As long as the FSDP communication is faster than the corresponding computation (which is the case for LLMs trained on up to hundreds of, say 512, GPUs), and there is no need to reduce (effective) per-GPU batch size to be below 1 (for reasons mentioned below in the TP section), 1D FSDP should be sufficient.

Existing ring-based implementations of NCCL collectives (all-gather, reduce-scatter) will incur a latency overhead which becomes severe at large scale (e.g. 512 GPUs). FSDP alone will become less efficient due to the collective latency increasing linearly with the world size, resulting in FSDP collectives that cannot be hidden by the computation any more. To further scale out, one needs to consider combining model parallelism solutions such as TP and PP.

#### 3.3.2 2D PARALLELISM: APPLY TP WITH FSDP

Model Parallelism (TP and PP) can help avoid the increased collective latency faced by scaling FSDP alone. TP can further lower the effective local batch size (to a minimum of  $\frac{1}{\text{TP degree}}$  when the local batch size is set to 1), as TP sharded models on multiple GPUs work jointly to process the same batch. This can be vital for reducing peak memory usage so that training could fit in GPU memory (e.g. due to large model size or sequence length), or for strong scaling with fixed desired global batch size (e.g. due to training efficiency considerations).

In addition, TP performs feature dimension sharding. This can bring more optimized matrix multiplication shapes for better FLOP utilization.

As TP introduces extra blocking collectives, in practice TP is only applied within a node that have fast interconnect (NVLink). AsyncTP could improve the performance by fully overlapping the communication, but could not scale to multi-node easily, so the TP degree is usually limited to 8. This means to scale beyond an ultra-large number of GPUs (e.g. more than 4192 GPUs), we need 3D parallelism that combines PP.

#### 3.3.3 3D PARALLELISM: APPLY PP WITH 2D PARALLELISM

Compared to other model parallelisms, PP requires less communication bandwidth by virtue of only transmitting activations and gradients between stages in a P2P manner. It is especially useful (1) to further reduce FSDP communication latency when the FSDP world size becomes large again that FSDP+TP still exposes FSDP collectives; or (2) to train with bandwidth-limited cluster.

We note that, the performance of PP, and in particular the “bubble” size, could vary by pipeline schedules being used and the microbatch size, assuming fixed global batch size and the world size.

## 4 DEMONSTRATING ADAPTABILITY AND EXTENSIBILITY

In this section, we demonstrate the adaptability and extensibility of LEGOSCALE by highlighting ongoing work and external contributions that showcase its ability to seamlessly integrate and compare new techniques, optimizations, and models.

### 4.1 ONGOING WORK: 4D PARALLELISM AND ZERO-BUBBLE PIPELINE SCHEDULES

LEGOSCALE’s modular and extensible architecture enables the seamless integration of new techniques and optimizations. For instance, ongoing work includes incorporating Context Parallel (Liu et al., 2023; Liu & Abbeel, 2024; NVIDIA, 2023) to enable 4D parallelism, and leveraging the `torch.distributed.pipelining` package to support zero-bubble schedules (Tang et al., 2024). This demonstrates LEGOSCALE’s ability to adapt to evolving machine learning landscapes.

### 4.2 EXTERNAL CONTRIBUTIONS: BUILDING AND EVALUATING CUSTOM INNOVATIONS

LEGOSCALE’s flexible architecture also empowers users to easily integrate and compare new innovations. By providing a modular and efficient test bed, LEGOSCALE enables users to rapidly benchmark new techniques, optimizations, and hardware on their training performance. This has already led to the refinement of a new production-grade dataloader, improvements in a new ZeRO implementation, advancements in an Adam-based optimizer, and the training of a top-tier diffusion model.

## 5 RELATED WORKS

With the rapidly growing significance of LLMs (Dubey et al., 2024; Achiam et al., 2023), there is substantial research and industry focus on improving infrastructure for training LLMs of various sizes. Since the very nature of these models are large, distributed training support becomes inevitable. Libraries like Megatron (Narayanan et al., 2021), DeepSpeed (Rasley et al., 2020), and PyTorch distributed (Pytorch native) (Paszke et al., 2019; Meta Platforms, Inc., 2024a) offer APIs to build distributed training workflows. NVIDIA NeMo (NVIDIA Corporation, 2024), built on Megatron-LM, offers a packaged solution for handling complex end-to-end model life-cycle from data curation to model deployment. Pytorch-native solutions like `torch tune` (Meta Platforms, Inc., 2024b) focus on fine-tuning LLMs in a simplified workflow. LEGOSCALE differs from these solutions by focusing on production grade pre-training using PyTorch-native APIs. The library is designed with elastic composability to accommodate the scale required to pre-train LLMs with minimal external dependencies. This lowers the bar to understand and extend pre-training, while offering features like async distributed checkpointing for building an end-to-end production workflow.

## 6 CONCLUSION

LEGOSCALE is a powerful and flexible framework for training LLMs. It offers composability, allowing users to combine various parallelism techniques (FSDP, TP, and PP), memory optimization methods (Float8 and activation checkpointing), and integration with PyTorch compiler to optimize training efficiency. LEGOSCALE is highly flexible, adaptable to evolving model architectures and hardware advancements, and features a modular design with multi-axis metrics that foster innovation and experimentation. LEGOSCALE also prioritizes interpretability, production-grade training, and PyTorch native capabilities. Additionally, it provides high-performance training with elastic scalability, comprehensive training recipes and guidelines, and expert guidance on selecting and combining distributed training techniques. As shown in the experiment sections, LEGOSCALE provides training accelerations ranging from 65.08% on Llama 3.1 8B at 128 GPU scale (1D), 12.59% on Llama 3.1 70B at 256 GPU scale (2D) to 30% on Llama 3.1 405B at 512 GPU scale (3D) over optimized baselines. With its robust features and high efficiency, LEGOSCALE is an ideal one-stop solution for challenging LLM training tasks.

## REFERENCES

- 540  
541  
542 Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany  
543 Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical re-  
544 port: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*,  
545 2024.
- 546 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
547 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
548 report. *arXiv preprint arXiv:2303.08774*, 2023.
- 549  
550 Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut,  
551 Johan Schalkwyk, Andrew M Dai, Anja Hauth, and Gemini Team. Gemini: a family of highly  
552 capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 553 Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky,  
554 Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will  
555 Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael  
556 Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael La-  
557 zos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan,  
558 Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting  
559 Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong  
560 Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. Py-  
561 torch 2: Faster machine learning through dynamic python bytecode transformation and graph  
562 compilation. In *Proceedings of the 29th ACM International Conference on Architectural Sup-  
563 port for Programming Languages and Operating Systems, Volume 2, ASPLOS '24*, pp. 929–947,  
564 New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703850. doi:  
565 10.1145/3620665.3640366. URL <https://doi.org/10.1145/3620665.3640366>.
- 566 James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal  
567 Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao  
568 Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/jax-ml/jax>.
- 569  
570 Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear  
571 memory cost, 2016. URL <https://arxiv.org/abs/1604.06174>.
- 572  
573 Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam  
574 Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm:  
575 Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):  
576 1–113, 2023.
- 577 Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding.  
578 *arXiv preprint arXiv:1810.04805*, 2018.
- 579  
580 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha  
581 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.  
582 *arXiv preprint arXiv:2407.21783*, 2024.
- 583 Assaf Eisenman, Kiran Kumar Matam, Steven Ingram, Dheevatsa Mudigere, Raghuraman Krish-  
584 namoorthi, Krishnakumar Nair, Misha Smelyanskiy, and Murali Annaram. Check-N-Run: a  
585 checkpointing system for training deep learning recommendation models. In *19th USENIX Sym-  
586 posium on Networked Systems Design and Implementation (NSDI 22)*, pp. 929–943, Renton, WA,  
587 April 2022. USENIX Association. ISBN 978-1-939133-27-4. URL <https://www.usenix.org/conference/nsdi22/presentation/eisenman>.
- 588  
589 Jiarui Fang and Shangchun Zhao. Usp: A unified sequence parallelism approach for long context  
590 generative ai, 2024. URL <https://arxiv.org/abs/2405.07719>.
- 591  
592 Tanmaey Gupta, Sanjeev Krishnan, Rituraj Kumar, Abhishek Vijeev, Bhargav Gulavani, Nipun  
593 Kwatra, Ramachandran Ramjee, and Muthian Sivathanu. Just-in-time checkpointing: Low cost  
error recovery from deep learning training failures. In *Proceedings of the Nineteenth European*

- 594 *Conference on Computer Systems*, EuroSys '24, pp. 1110–1125, New York, NY, USA, 2024. As-  
595 sociation for Computing Machinery. ISBN 9798400704376. doi: 10.1145/3627703.3650085.  
596 URL <https://doi.org/10.1145/3627703.3650085>.  
597
- 598 Horace He and Shangdi Yu. Transcending runtime-memory tradeoffs in checkpointing by being  
599 fusion aware. *Proceedings of Machine Learning and Systems*, 5:414–427, 2023.
- 600 Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Mia Xu Chen, Dehao Chen, Hy-  
601 oukJoong Lee, Jiquan Ngiam, Quoc V. Le, Yonghui Wu, and Zhifeng Chen. *GPipe: efficient*  
602 *training of giant neural networks using pipeline parallelism*. Curran Associates Inc., Red Hook,  
603 NY, USA, 2019.
- 604 Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bam-  
605 ford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al.  
606 Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.  
607
- 608 Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael An-  
609 dersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recom-  
610 putation in large transformer models. In D. Song, M. Carbin, and T. Chen  
611 (eds.), *Proceedings of Machine Learning and Systems*, volume 5, pp. 341–353. Cu-  
612 ran, 2023. URL [https://proceedings.mlsys.org/paper\\_files/paper/2023/](https://proceedings.mlsys.org/paper_files/paper/2023/file/80083951326cf5b35e5100260d64ed81-Paper-mlsys2023.pdf)  
613 [file/80083951326cf5b35e5100260d64ed81-Paper-mlsys2023.pdf](https://proceedings.mlsys.org/paper_files/paper/2023/file/80083951326cf5b35e5100260d64ed81-Paper-mlsys2023.pdf).
- 614 Jianhui Li, Zhennan Qin, Yijie Mei, Jingze Cui, Yunfei Song, Ciyong Chen, Yifei Zhang, Longsheng  
615 Du, Xianhang Cheng, Baihui Jin, Yan Zhang, Jason Ye, Eric Lin, and Dan Lavery. onednn  
616 graph compiler: A hybrid approach for high-performance deep learning compilation. In *2024*  
617 *IEEE/ACM International Symposium on Code Generation and Optimization (CGO)*, pp. 460–  
618 470, 2024. doi: 10.1109/CGO57630.2024.10444871.
- 619 Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff  
620 Smith, Brian Vaughan, Pritam Damania, et al. Pytorch distributed: Experiences on accelerating  
621 data parallel training. *arXiv preprint arXiv:2006.15704*, 2020.
- 622 Hao Liu and Pieter Abbeel. Blockwise parallel transformers for large context models. *Advances in*  
623 *Neural Information Processing Systems*, 36, 2024.  
624
- 625 Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-  
626 infinite context. *arXiv preprint arXiv:2310.01889*, 2023.
- 627 Yinhan Liu, Myle Ott, and Naman Goyal. Jingfei du, mandar joshi, danqi chen, omer levy, mike  
628 lewis, luke zettlemoyer, and veselin stoyanov. 2019. roberta: A robustly optimized bert pretraining  
629 approach. *arXiv preprint arXiv:1907.11692*, 1(3.1):3–3, 2019.
- 630 Avinash Maurya, Robert Underwood, M. Mustafa Rafique, Franck Cappello, and Bogdan Nicolae.  
631 Datatypes-llm: Lazy asynchronous checkpointing for large language models. In *Proceedings*  
632 *of the 33rd International Symposium on High-Performance Parallel and Distributed Comput-*  
633 *ing*, HPDC '24, pp. 227–239, New York, NY, USA, 2024. Association for Computing Machin-  
634 ery. ISBN 9798400704130. doi: 10.1145/3625549.3658685. URL [https://doi.org/10.](https://doi.org/10.1145/3625549.3658685)  
635 [1145/3625549.3658685](https://doi.org/10.1145/3625549.3658685).  
636
- 637 Meta Platforms, Inc. PyTorch Distributed, 2024a. URL [https://pytorch.org/docs/](https://pytorch.org/docs/stable/distributed.html)  
638 [stable/distributed.html](https://pytorch.org/docs/stable/distributed.html). Accessed: 2023-09-26.
- 639 Meta Platforms, Inc. PyTorch TorchTune, 2024b. URL [https://pytorch.org/](https://pytorch.org/torchtune/stable/overview.html)  
640 [torchtune/stable/overview.html](https://pytorch.org/torchtune/stable/overview.html). Accessed: 2024-09-26.
- 641 Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia,  
642 Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed  
643 precision training, 2018. URL <https://arxiv.org/abs/1710.03740>.  
644
- 645 Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisen-  
646 thwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi,  
647 Stuart Oberman, Mohammad Shoeybi, Michael Siu, and Hao Wu. Fp8 formats for deep learning,  
2022. URL <https://arxiv.org/abs/2209.05433>.

- 648 Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R. Devanur, Gre-  
649 gory R. Ganger, Phillip B. Gibbons, and Matei Zaharia. Pipedream: generalized pipeline par-  
650 allelism for dnn training. In *Proceedings of the 27th ACM Symposium on Operating Systems*  
651 *Principles*, SOSP '19, pp. 1–15, New York, NY, USA, 2019. Association for Computing Machin-  
652 ery. ISBN 9781450368735. doi: 10.1145/3341301.3359646. URL [https://doi.org/10.](https://doi.org/10.1145/3341301.3359646)  
653 [1145/3341301.3359646](https://doi.org/10.1145/3341301.3359646).
- 654 Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vi-  
655 jay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar  
656 Phanishayee, and Matei Zaharia. Efficient large-scale language model training on gpu clus-  
657 ters using megatron-lm. In *Proceedings of the International Conference for High Performance*  
658 *Computing, Networking, Storage and Analysis*, SC '21, New York, NY, USA, 2021. Associa-  
659 tion for Computing Machinery. ISBN 9781450384421. doi: 10.1145/3458817.3476209. URL  
660 <https://doi.org/10.1145/3458817.3476209>.
- 661 NVIDIA. Megatron Core API Guide: Context Parallel, 2023. URL [https://docs.nvidia.](https://docs.nvidia.com/megatron-core/developer-guide/latest/api-guide/context_parallel.html)  
662 [com/megatron-core/developer-guide/latest/api-guide/context\\_](https://docs.nvidia.com/megatron-core/developer-guide/latest/api-guide/context_parallel.html)  
663 [parallel.html](https://docs.nvidia.com/megatron-core/developer-guide/latest/api-guide/context_parallel.html). Accessed: 2023-09-25.
- 664 NVIDIA Corporation. NVIDIA Nemo, 2024. URL [https://www.nvidia.com/en-us/](https://www.nvidia.com/en-us/ai-data-science/products/nemo/)  
665 [ai-data-science/products/nemo/](https://www.nvidia.com/en-us/ai-data-science/products/nemo/). Accessed: 2024-09-26.
- 666 Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor  
667 Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Ed-  
668 ward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner,  
669 Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep*  
670 *learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- 671 PyTorch Community. Pytorch dtensor rfc, 2023a. URL [https://github.com/pytorch/](https://github.com/pytorch/pytorch/issues/88838)  
672 [pytorch/issues/88838](https://github.com/pytorch/pytorch/issues/88838). GitHub Issue.
- 673 PyTorch Community. Float8 in pytorch 1.x, 2023b. URL [https://dev-discuss.pytorch.](https://dev-discuss.pytorch.org/t/float8-in-pytorch-1-x/1815)  
674 [org/t/float8-in-pytorch-1-x/1815](https://dev-discuss.pytorch.org/t/float8-in-pytorch-1-x/1815). PyTorch Discussion Thread.
- 675 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language  
676 models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 677 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi  
678 Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text  
679 transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.
- 680 Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: memory optimizations  
681 toward training trillion parameter models. SC '20. IEEE Press, 2020. ISBN 9781728199986.
- 682 Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System op-  
683 timizations enable training deep learning models with over 100 billion parameters. KDD '20,  
684 pp. 3505–3506, New York, NY, USA, 2020. Association for Computing Machinery. ISBN  
685 9781450379984. doi: 10.1145/3394486.3406703. URL [https://doi.org/10.1145/](https://doi.org/10.1145/3394486.3406703)  
686 [3394486.3406703](https://doi.org/10.1145/3394486.3406703).
- 687 Ding Tang, Lijuan Jiang, Jiecheng Zhou, Minxi Jin, Hengjie Li, Xingcheng Zhang, Zhilin Pei, and  
688 Jidong Zhai. Zeropp: Unleashing exceptional parallelism efficiency through tensor-parallelism-  
689 free methodology, 2024. URL <https://arxiv.org/abs/2402.03791>.
- 690 Borui Wan, Mingji Han, Yiyao Sheng, Zhichao Lai, Mofan Zhang, Junda Zhang, Yanghua Peng,  
691 Haibin Lin, Xin Liu, and Chuan Wu. Bytecheckpoint: A unified checkpointing system for llm  
692 development, 2024. URL <https://arxiv.org/abs/2407.20143>.
- 693 Shibo Wang, Jinliang Wei, Amit Sabne, Andy Davis, Berkin Ilbeyi, Blake Hechtman, Dehao Chen,  
694 Karthik Srinivasa Murthy, Marcello Maggioni, Qiao Zhang, et al. Overlap communication with  
695 dependent computation via decomposition in large deep learning models. In *Proceedings of the*  
696 *28th ACM International Conference on Architectural Support for Programming Languages and*  
697 *Operating Systems, Volume 1*, pp. 93–106, 2022.

Zhuang Wang, Zhen Jia, Shuai Zheng, Zhen Zhang, Xinwei Fu, T. S. Eugene Ng, and Yida Wang. Gemini: Fast failure recovery in distributed training with in-memory checkpoints. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23*, pp. 364–381, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613145. URL <https://doi.org/10.1145/3600006.3613145>.

Cody Hao Yu, Haozheng Fan, Guangtai Huang, Zhen Jia, Yizhi Liu, Jie Wang, Zach Zheng, Yuan Zhou, Haichen Shen, Junru Shao, Mu Li, and Yida Wang. Raf: Holistic compilation for deep learning model training, 2023. URL <https://arxiv.org/abs/2303.04759>.

Buyun Zhang, Liang Luo, Xi Liu, Jay Li, Zeliang Chen, Weilin Zhang, Xiaohan Wei, Yuchen Hao, Michael Tsang, Wenjun Wang, Yang Liu, Huayu Li, Yasmine Badr, Jongsoo Park, Jiyan Yang, Dheevatsa Mudigere, and Ellie Wen. Dhen: A deep and hierarchical ensemble network for large-scale click-through rate prediction, 2022. URL <https://arxiv.org/abs/2203.11014>.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860, aug 2023. ISSN 2150-8097. doi: 10.14778/3611540.3611569. URL <https://doi.org/10.14778/3611540.3611569>.

## A COMPOSABLE 3D PARALLELISM WALKTHROUGH

We have discussed the scaling with LEGOSCALE 3D parallelism and the motivations to apply different parallelisms to scale training to thousands of GPUs. In this section we will walk through the 3D parallelism code in LEGOSCALE.

The first step is to create an instance of the model (e.g. the Transformer for Llama models) on the meta device. We then apply PP by splitting the model into multiple PP stages according to the `pipeline_parallel_split_points` config. Note that for PP with looped schedules, we may obtain multiple `model_parts` from PP splitting, where each item in `model_parts` is one stage-model-chunk. Next we apply SPMD-style distributed training techniques including TP, activation checkpointing, `torch.compile`, FSDP, and mixed precision training for each model part, before actually initializing the sharded model on GPU.

---

```

734 # meta init
735 with torch.device("meta"):
736     model = model_cls.from_model_args(model_config)
737
738 # apply PP
739 pp_schedule, model_parts = models_pipelining_fns[model_name](
740     model, pp_mesh, parallel_dims, job_config, device, model_config,
741     loss_fn
742 )
743
744 for m in model_parts:
745     # apply SPMD-style distributed training techniques
746     models_parallelize_fns[model_name](m, world_mesh, parallel_dims,
747         job_config)
748     # move sharded model to GPU and initialize weights via DTensor
749     m.to_empty(device="cuda")
750     m.init_weights()

```

---

To apply PP to the model, we run the following code at the high level. `pipeline_llama_manual_split` splits the model into multiple stages according to the manually given `pipeline_parallel_split_points` config, by removing the unused model components from a complete model (on the meta device). Then `build_pipeline_schedule` make the pipeline schedule with various options from `torch.distributed.pipelining`, including 1F1B (Narayanan et al., 2019), GPipe (Huang et al., 2019), interleaved 1F1B (Narayanan et al., 2021), etc. instructed by the `pipeline_parallel_schedule` config.

---

```

756
757 stages, models = pipeline_llama_manual_split(
758     model, pp_mesh, parallel_dims, job_config, device, model_config
759 )
760 pp_schedule = build_pipeline_schedule(job_config, stages, loss_fn)
761 return pp_schedule, models

```

---

TP and FSDP are applied in the SPMD-style `models_parallelize_fns` function. To apply TP, we utilize the DTensor `parallelize_module` API, by providing a TP “plan” as the instruction of how model parameters should be sharded. In the example below, we showcase the (incomplete) code for sharding the repeated `TransformerBlock`.

---

```

767 for layer_id, transformer_block in model.layers.items():
768     layer_tp_plan = {
769         "attention_norm": SequenceParallel(),
770         "attention": PrepareModuleInput(
771             input_layouts=(Shard(1), None),
772             desired_input_layouts=(Replicate(), None),
773         ),
774         "attention.wq": ColwiseParallel(),
775         ...
776     }
777     parallelize_module(
778         module=transformer_block,
779         device_mesh=tp_mesh,
780         parallelize_plan=layer_tp_plan,
781     )

```

---

Finally, we apply the FSDP by wrapping each individual `TransformerBlock` and then the whole model. Note that the FSDP2 implementation in PyTorch comes with mixed precision training support. By default, we use `torch.bfloat16` on parameters all-gather and activation computations, and use `torch.float32` on gradient reduce-scatter communication and optimizer updates.

---

```

785 mp_policy = MixedPrecisionPolicy(param_dtype, reduce_dtype)
786 fsdp_config = {"mesh": dp_mesh, "mp_policy": mp_policy}
787
788 for layer_id, transformer_block in model.layers.items():
789     # As an optimization, do not reshard_after_forward for the last
790     # TransformerBlock since FSDP would prefetch it immediately
791     reshard_after_forward = int(layer_id) < len(model.layers) - 1
792     fully_shard(
793         transformer_block,
794         **fsdp_config,
795         reshard_after_forward=reshard_after_forward,
796     )
797     fully_shard(model, **fsdp_config)

```

---

## 800 B SUPPLEMENTARY MATERIALS

### 802 B.1 FULLY SHARDED DATA PARALLEL

804 FSDP2 makes improvements over the original FSDP1 FlatParameter grouping. Specifically, parameters are now represented as DTensors sharded on the tensor dimension 0. This provides better composability with model parallelism techniques and other features that requires the manipulation of individual parameters, allowing sharded state dict to be represented by DTensor without any communication, and provides for a simpler meta-device initialization flow via DTensor. For example, FSDP2 unlocks finer grained tensor level quantization, especially Float8 tensor quantization, which we will showcase in the results section.

810 As part of the rewrite from FSDP1 to FSDP2, FSDP2 implements an improved memory management  
811 system by avoiding the use of record stream. This enables deterministic memory release, and as a  
812 result provides lower memory requirements per GPU relative to FSDP1. For example on Llama 2  
813 7B, FSDP2 records an average of 7% lower GPU memory versus FSDP1.

814 In addition, by writing efficient kernels to perform multi-tensor allgather and reduce scatter, FSDP2  
815 shows on-par performance compare to FSDP1, an there are slight performance gains from FSDP2 -  
816 using the Llama 2 7B, FSDP2 shows an average gain of 1.5% faster throughput.

817 The performance gains are the result of employing two small performance improvements. First,  
818 only a single division kernel is run for the FP32 reduce scatter (pre-dividing the local FP32 reduce-  
819 scatter gradient by world size, instead of a two step pre and post divide by square root of world size).  
820 Secondly, in LEGOSCALE, FSDP2 is integrated with a default of not sharding the final block in a  
821 transformer layer during the forward pass, since it will be immediately re-gathered at the start of the  
822 backward pass. Thus we can skip a round of communications delay.

823 **Usage:** LEGOSCALE has fully integrated FSDP2 as the default parallelism when training, and the  
824 `data_parallel_shard_degree` is the controlling dimension in the command line or TOML  
825 file. Note that for ease of use, leaving `data_parallel_shard_degree` as -1, which is the  
826 default, means to simply use all GPU's available (i.e. no need to spec your actual world size).  
827

## 828

### 829 B.2 HYBRID SHARDED DATA PARALLEL

### 830

831 Hybrid Sharded Data Parallel (HSDP) is an extension of FSDP (Zhang et al., 2022), which enables a  
832 larger total world size to be used. In FSDP, all devices are part of a single global group across which  
833 all communications are enabled. However, at some point, adding more computation is offset by the  
834 increasing communication overhead due to adding more participants which require equal commu-  
835 nication participation. This is due to the fact that the latency of collective communications have a  
836 direct correlation with the total number of participants. At this saturation point, FSDP throughput  
837 will effectively flat-line even as more computation is added. HSDP obviates this to some degree  
838 by creating smaller sharding groups (islands) within the original global group (ocean), where each  
839 sharding group runs FSDP amongst itself, and gradients are synced across sharding groups at set  
840 frequency during the backward pass to ensure a global gradient is maintained. This ensures speedy  
841 communications as the total participant communication size is now a fraction of the original world  
842 size, and the only global communication is for the gradient all-reduce between the sharding groups.  
843 By using sharding groups, we have seen that HSDP can extend the total world size by 3-6x rela-  
844 tive to FSDP's communication saturation point (this will vary, depending on the speed of network  
interconnects).

845 LEGOSCALE makes it easy to run HSDP with two user configurable settings for sharding group size  
846 and replication group size, from the command line or TOML file.

847 **Usage:** HSDP is enabled in LEGOSCALE by modifying the previously mentioned knob  
848 `data_parallel_shard_degree` to control the sharding group size. This is effectively the  
849 `gpu_group` count that will run FSDP sharding among its corresponding group members. From there,  
850 we must spec the `data_parallel_replicate_degree`, which controls how many sharding  
851 groups we are creating. The product of both replicate and shard degree must add up to the total  
852 world size. Example - on a 128 GPU cluster, we may find that sharding over 16 gpus would be  
853 enough for the model size. Therefore, we set the `data_parallel_shard_degree` to be 16,  
854 and the `data_parallel_replicate_degree` be 8 correspondingly, meaning we will have 8  
855 groups of 16 GPUs to fill out the total world size of 128.

### 856

### 857

### 858 B.3 TENSOR PARALLEL

### 859

860 TP partitions the attention and feed forward network (MLP) modules of a transformer layer across  
861 multiple devices, where the number of devices used is the TP degree. This allows for multiple GPU's  
862 to cooperatively process a transformer layer that would otherwise exceed a single GPU's ability, at  
863 the cost of adding `all-reduce/all-gather/reduce-scatter` operations to synchronize intermediates.



864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917

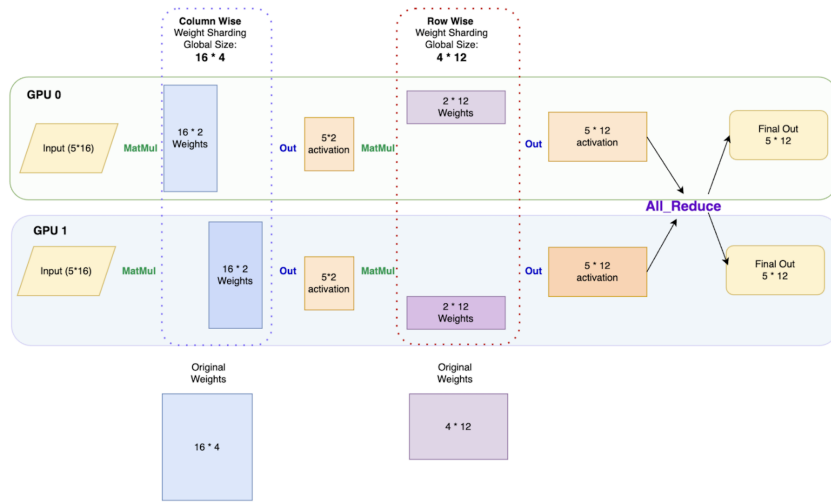


Figure 3: Tensor Parallel in detail (2 GPUs, data moves from left to right).

Due to the additional collectives introduced by TP, it needs to happen on a fast network (i.e NVLink). When training LLMs, TP is usually combined with FSDP, where TP shards within nodes and FSDP shards across nodes to create the 2D hierarchical sharding on different DeviceMesh dimensions.

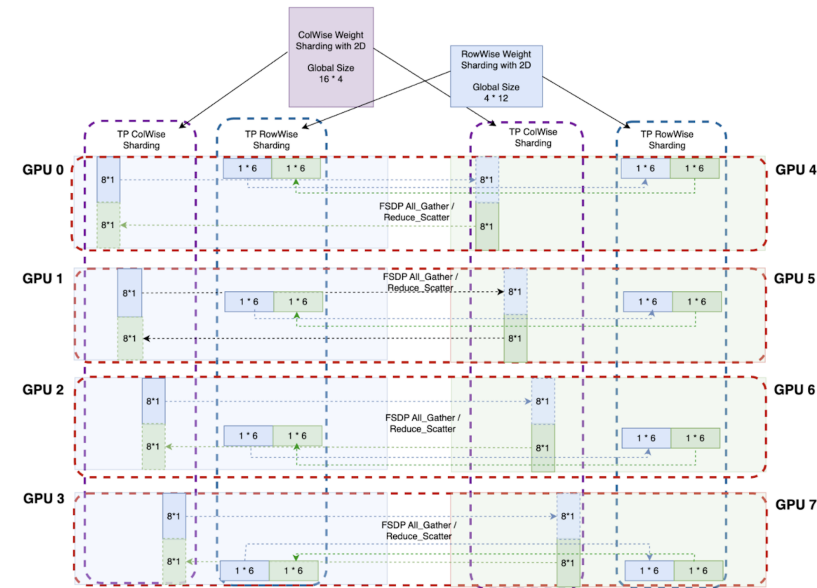


Figure 4: FSDP2 + Tensor Parallel (TP degree 4) sharding layout, with 2 nodes of 4 GPUs.

**Usage:** Because of the synergistic relationship between TP and SP, LEGOSCALE natively bundles these two together and they are jointly controlled by the TP degree setting in the command line or the TOML entry of `tensor_parallel_degree`. Setting this to 2 for example would mean that 2 GPUs within the node will share the computational load for each transformer layers attention and MLP modules via TP, and normalization/dropout layers via Sequence Parallel. Loss Parallel is implemented via a context manager as it needs to control the loss computation outside of the model's forward computation. It can be enabled via `enable_loss_parallel`.

#### B.4 PIPELINE PARALLEL

We expose several parameters to configure PP. `pipeline_parallel_degree` controls the number of ranks participating in PP. `pipeline_parallel_split_points` accepts a list of strings, representing layer fully-qualified-names before which a split will be performed. Thus, the total number of pipeline stages  $V$  will be determined by the length of this list. `pipeline_parallel_schedule` accepts the name of the schedule to be used. If the schedule is multi-stage, there should be  $V > 1$  stages assigned to each pipeline rank, otherwise  $V == 1$ . `pipeline_parallel_microbatches` controls the number of microbatches to split a data batch into.

#### B.5 ACTIVATION CHECKPOINTING

LEGOSCALE offers two types of Selective Activation Checkpointing which allow for a more nuanced tradeoff between memory and recomputation. Specifically, we offer the option to selectively checkpoint “per layer” or “per operation”. The goal for per operation is to free memory used by operations that are faster to recompute and save intermediates (memory) for operations that are slower to recompute and thus deliver a more effective throughput/memory trade-off.

**Usage:** AC is enabled via a two-line setting in the command line or TOML file. Specifically, `mode` can be either `none`, `selective`, or `full`. When `selective` is set, then the next config of `selective_ac_type` is used which can be either a positive integer to enable selective layer checkpointing, or `op` to enable selective operation checkpointing. Per layer takes an integer input to guide the checkpointing policy, where 1 = checkpoint every layer (same as full), 2 = checkpoint every other layer, 3 = checkpoint every third layer, etc. Per operation is driven by the `_save_list` policy in `parallelize_llama.py` which flags high arithmetic intensity operations such as matmul (matrix multiplication) and SPDA (Scaled Dot Product Attention) for saving the intermediate results, while allowing other lower intensity operations to be recomputed. Note that for balancing total throughput, only every other matmul is flagged for saving.

#### B.6 ASYNCTP

The `SymmetricMemory` collectives used in AsyncTP are faster than standard NCCL collectives and operate by having each GPU allocate an identical memory buffer in order to provide direct P2P access. `SymmetricMemory` relies on having NVSwitch within the node, and is thus generally only available for H100 or newer GPUs.

**Usage:** AsyncTP is enabled within the experimental section of the LEGOSCALE TOML config file and turned on or off via the `enable_async_tensor_parallel` boolean setting.

#### B.7 CUSTOMIZING FSDP2 MIXED PRECISION IN LEGOSCALE

Mixed Precision is controlled by the `MixedPrecisionPolicy` class in the `apply_fsdp` function, which is then customized with `param_dtype` as BF16, and `reduce_dtype` defaulting to FP32 by default in LEGOSCALE. The `reduce_dtype` in FP32 means that the reduce-scatter in the backwards pass for gradient computation will take place in FP32 to help maximize both stability and precision of the gradient updates.