

Deep Learning System Design



Engineering and Service Architectures

Han Cheol Moon
School of Computer Science and Engineering
Nanyang Technological University

Singapore
hancheol001@e.ntu.edu.sg
August 25, 2025

Contents

I	Introduction	1
1	Introduction	2
1.1	Operations challenges with LLMs	2
1.2	LLMOps Essentials	2
1.3	LLM Operations Infrastructures	3
2	Preliminaries	5
2.1	Complexity of Matrix Multiplication	5
3	Data Engineering for LLMs	7
3.1	Models and the Foundation	7
3.1.1	Evaluating LLMs	7
II	Parallelism	8
4	Data Parallelism	9
4.1	Data Parallel	9
4.2	Distributed Data Parallel	9
4.2.1	Concepts and Terminology	10
4.2.2	How DDP Works Under the Hood	11
5	Pipeline Parallelism	12
5.1	Introduction	12
5.1.1	Illustration of the Pipeline	12

<i>CONTENTS</i>	2
5.1.2 Pipeline Bubbles	14
5.1.3 Combining Pipeline Parallelism with Other Forms of Parallelism	14
5.2 1F1B	15
5.2.1 Non-interleaved Schedule	15
5.2.2 Interleaved Schedule	15
5.3 Zero Bubble	15
6 Tensor Parallelism	16
6.1 Introduction	16
7 N-Dim Parallelism	19
8 DualPipe	20
8.1 Introduction	20
8.1.1 All-to-All vs Point-to-Point	20
8.2 DualPipe	20
III Transformers	22
8.3 Flash Attention	23
9 Tokenization	24
IV Compression	25
10 Model Compression	26
10.1 Introduction	26

Part I

Introduction

Chapter 1

Introduction

1.1 Operations challenges with LLMs

- **Long download times** (*e.g.*, Bloom LLM is 330GB).
- **Longer deploy times** (*e.g.*, Bloom takes 30 ~ 45 mins to load the model into GPU).
- Along with increases in model size often come increases in **inference latency**.
- **Managing GPUs**
- **Peculiarities of text data**: unlike other fields, texts have ambiguities.
- **Token limits for a model** create bottlenecks
- **Hallucinations cause confusion**
- **Bias and ethical considerations**
- **Security concerns**
- **Controlling costs**: *e.g.*, GPUs, infra, storage, operational costs like energy consumption during both training and inference.

1.2 LLMOps Essentials

- **Compression** is the practice of making models smaller.
 - **Quantizing** is the process of reducing precision in preference of lowering the memory requirements.
 - **Pruning** is the process of weeding out and removing any parts of the model we deem unworthy.
 - **Knowledge distillation** takes the large LLM and train a smaller language model to copy it.
 - **Low-rank approximation** is a trick to simplify large matrices or tensors to find a lower dimensional representation.

- **Mixture of Experts** (MoE) is a technique where we replace the feed-forward (FF) layers in a transformer with MoE layers instead. FF layers are notorious for being parameter-dense and computationally intensive, so replacing them with something better can often have a large effect. MoEs are a group of sparsely activated models. They differ from ensemble techniques in that typically only one or a few expert models will be run, rather than combining results from all models. The sparsity is often induced by a *gate mechanism* that learns which experts to use and/or a router mechanism that determines which experts should even be consulted.
- **Distributed computing** is a technique used in DL to parallelize and speed up large, complex neural networks by dividing the workload across multiple devices or nodes in a cluster. This approach significantly reduces training and inference times by enabling concurrent computation, data parallelism, and model parallelism.
 - **Data parallelism**: splitting up the data and running them through multiple copies of the model or pipeline.
 - **Tensor parallelism**: This approach takes advantage of matrix multiplication properties of split up the activations across multiple processors, running the data through and then combining them on the other side of the processors.
 - **Pipeline parallelism**: This creates a pipeline, as input data will go to the first GPU, process, then transfer to the next GPU, and so on until it's run through the entire model.
 - **3D parallelism**: We want to take advantages of all three parallelism practices as they can all be run together. This is known as 3D parallelism, which combines data, tensor and pipeline parallelism (DP+TP+PP) together. Since each technique and thus dimension will require at least two GPUs to run 3D parallelism, we will need at least eight GPUs to get started.

1.3 LLM Operations Infrastructures

- Data infrastructure is the foundation of DataOps. (*e.g.*, Airflow, Prefect, and Mage)
- Experiment trackers (*e.g.*, MLFlow and Weights & Biases)
- Model registry
- Feature stores (*e.g.*, Feast) is a centralized system for managing, storing, and serving features (the inputs to machine learning models) in a consistent and reliable way. It sits between your raw data sources and your ML models, making it easier to build, deploy, and maintain production ML systems. Why do we need it? In ML, the same features are used in two places:
 - Training – when building the model.
 - Serving/Inference – when the model is deployed to make predictions.
- Vector databases (*e.g.*, Qdrant, Pinecone, and Milvus) are specialized databases that store vectors along with some metadata around the vector, which makes them great for storing embeddings. The power of vector databases isn't in their storage but in the way that they search through the data.

- Monitoring systems (*e.g.*, whylogs and ¹): ML models are often fail silently (*e.g.*, data drift ¹).
- GPU-enabled workstations (*e.g.*, H100 provides 80GB and NVL ²)
 - If you aren't sure which GPU you need to run which model, multiply the number of billions of parameters by two, since most models at inference will default to run at half precision, FP16 or BF16, which means we need at least 2 bytes for every parameter. You will need a little extra as well for embedding model, which will be about another gigabyte, and more for the actual tokens. One token is about 1MB. For 16 batches of this size, you will need an extra 8GB of space.
 - For training you will need a lot more space (*e.g.*, full precision, optimizer tensors and gradients). Roughly you need 16 bytes for every parameter, so to train a 7B parameter model, you will need 112GB of memory.
- Deployment service (*e.g.*, NVIDIA Triton Inference Service, MLServer, Seldon, BentoML)

¹Data drift in Machine Learning (ML) systems refers to changes in the statistical properties of input data (or the relationship between inputs and outputs) over time, which can degrade model performance if not monitored and addressed.

²NVIDIA Link is NVIDIA's high-speed interconnect technology. It allows GPUs (and sometimes CPUs) to communicate with each other much faster than PCIe. In multi-GPU systems (like deep learning servers, HPC, or AI supercomputers), GPUs need to share data frequently (*e.g.*, gradients and parameters). If they only use PCIe, communication is slower, creating a bottleneck.

Chapter 2

Preliminaries

2.1 Complexity of Matrix Multiplication

Matrix multiplication is a fundamental operation in many computational tasks, including neural networks. The complexity of multiplying two matrices depends on their dimensions. Let's dive into the specifics.

- Let A be a matrix of size $m \times k$.
- Let B be a matrix of size $k \times n$.
- The result C will be a matrix of size $m \times n$.

Standard Matrix Multiplication: For each element c_{ij} in the resulting matrix C :

$$c_{ij} = \sum_{l=1}^k a_{il} \cdot b_{lj}$$

This involves:

- Multiplications: k multiplications for each element c_{ij} .
- Additions: $k - 1$ additions for each element c_{ij} .

Complexity

- The total number of elements in C is $m \times n$.
- Therefore, the total number of multiplications is $m \times n \times k$.
- The total number of additions is $m \times n \times (k - 1)$.

Thus, the total complexity is $O(m \times n \times k)$.

Even though there are several advanced methods, the standard $O(m \times n \times k)$ complexity is often used in practice, due to the simplicity and efficiency of implementation on modern hardware. Optimized libraries (like BLAS, cuBLAS for GPUs) leverage hardware-specific optimizations to improve practical performance.

Chapter 3

Data Engineering for LLMs

Data engineering is the development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information that supports downstream use cases, such as analysis and machine learning.

There isn't more valuable asset than your data. All successful AI and ML initiatives are built on a good data engineering foundation. It's important then that we acquire, clean, and curate our data.

3.1 Models and the Foundation

The most important dataset you will need to collect when training is the model weights of a pretrained model.

3.1.1 Evaluating LLMs

When evaluating a model, you will need two things: *(i)* a *metric* and *(ii)* a *dataset*.

Metrics

- ROUGE
- BLEU
- BPC: The bits per character (BPC) evaluation is an example of an entropy-based evaluation for language models.

Industry benchmarks

- GLUE
- SuperGLUE

Part II

Parallelism

Chapter 4

Data Parallelism

4.1 Data Parallel

The first step of the typical training loop for deep learning models is to split a dataset into batches so that we can feed them into the model and compute gradients corresponding to them. As the model size grows up, we couldn't fit the model into a single GPU. The *data parallelism* tries to tackle the issue by clone the model across multiple GPUs so that each GPU can take a small portion of the batches for each iteration. Data Parallel (sometimes referred to as “single-node data parallel”) is typically used when you have **multiple GPUs on a single machine**.

Let's say the batch size is 10 and we have 5 GPUs. Then, each GPU takes 2 batches and calculate gradients by on its own. The calculated gradients are then synchronized across the GPUs pretending they are computed on a single GPU. Finally, the synchronized gradient information is going to be distributed to them.

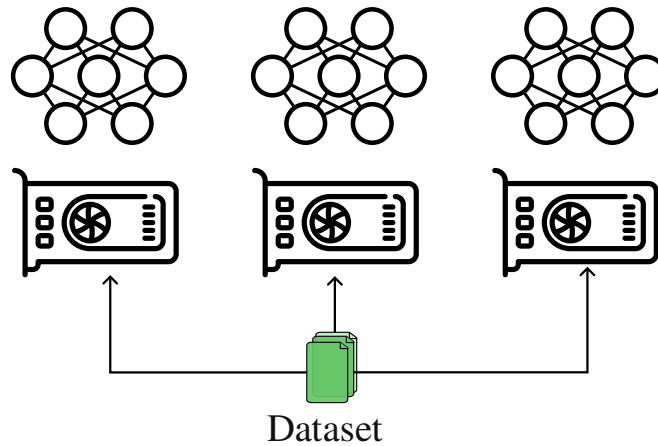
There are some important things to mention:

1. One process (or master thread) becomes a bottleneck for gradient aggregation and parameter updates.
2. As you increase the number of GPUs, or try to involve multiple machines, communication overhead grows significantly and can slow down training.
3. Each GPU holds a copy of the entire model, which can be large.

4.2 Distributed Data Parallel

To alleviate such issues, we can adopt an approach called *Distributed Data Parallel* (DDP), which is designed to scale training across many GPUs, potentially across multiple machines (nodes). Modern deep learning frameworks (like PyTorch `torch.nn.parallel.DistributedDataParallel`) typically recommend DDP as the best practice for multi-GPU/multi-node training due to better performance and scalability. During backpropagation, gradients are shared among GPUs through efficient communication primitives, resulting in synchronized model parameters across all GPUs.

Key benefits:



- Scalability: You can increase the number of GPUs (and even add more machines) to handle large datasets and bigger models.
- Performance: DDP typically provides better performance than older methods like `NN.DATAPARALLEL` (in PyTorch) because it uses *all-reduce* and eliminates the single “master” bottleneck.
- Flexibility: You can combine DDP with other parallelization strategies (*e.g.*, model parallel, sharded data parallel, pipeline parallel) if needed.

4.2.1 Concepts and Terminology

All-Reduce is a collective communication operation commonly used in distributed computing (especially in high-performance computing and deep learning). In simple terms:

- Each process (or GPU) starts with its own data (*e.g.*, local gradients).
- These data are combined (usually via a reduction operation like sum, mean, min, or max) across all processes.
- The result of that reduction (*e.g.*, the summed gradients) is then shared back so that every process receives the same reduced value.
- Hence the name: “all” (everyone gets the result) + “reduce” (combine data).

Basic Terms:

- World Size: The total number of processes engaged in the distributed job. Often, we run one process per GPU, so world size is the number of GPUs.
- Rank: A unique integer ID assigned to each process. Ranks typically range from 0 to `world_size - 1`. Rank 0 is often referred to as the “leader” or “master” process, but in DDP, every process does roughly the same work.
- Local Rank: When multiple GPUs reside on a single node, local rank identifies which GPU a specific process is mapped to on that local machine (*e.g.*, 0 for the first GPU, 1 for the second, etc.).

- Backend: The communication backend used for synchronization (*e.g.*, nccl). For GPU training, NCCL is typically recommended because it's optimized for high-performance GPU-to-GPU communication.
- Initialization Method: Describes how processes connect with each other (*e.g.*, a TCP store, a file-based store). This allows all processes to know who's who in the cluster.

4.2.2 How DDP Works Under the Hood

1. Process Per GPU: Each GPU runs the same script in its own process.
2. Data Subset: A DistributedSampler ensures that each process sees a unique subset of data. This prevents overlap in data usage among GPUs.
3. Full Model Copy: Each GPU has a full replica of the model in memory.
 - For massive models, consider *Sharded DDP* (*e.g.*, PyTorch's FSDP or DeepSpeed ZeRO) to split parameters across GPUs.
4. All-Reduce Gradient Sync: After backprop, gradients are summed (or averaged) across processes with an all-reduce operation. This keeps all models in sync.

Chapter 5

Pipeline Parallelism

5.1 Introduction

The basic idea of the data parallel is to distribute the model across GPUs. However, if the model size is bigger than the VRAM of GPU, the model wouldn't fit in a single GPU. To resolve the issue, we have to split the model across GPUs. For instance, we can put the half of the model into the first GPU and the remaining half into the second GPU. This approach is often called *model parallelism*. Let's closely look at one of the model parallelism approaches, called *pipeline parallelism*.

Pipeline Parallelism is a strategy for distributing large deep learning models across multiple devices (GPUs) by splitting the model layers into sequential stages. Rather than replicating the entire model on each GPU or sharding the parameters themselves, pipeline parallelism assigns a subset of layers to each device in a pipeline-like fashion. This technique is especially helpful when:

- The model is too large to fit on a single GPU, but it can be split into chunks (layers/stages).
- You want to keep multiple GPUs actively working on different portions (stages) of the forward and backward pass concurrently.

5.1.1 Illustration of the Pipeline

In pipeline parallelism, the model is divided into N sub-networks, and each sub-network is placed on a different GPU (or sometimes on multiple GPUs if you have many layers). Think of it like an assembly line:

- Sub-Network 1: Layers $1 - k$
- Sub-Network 2: Layers $(k + 1) - m$
- Sub-Network 3: Layers $(m + 1) - \dots$
- and so on.

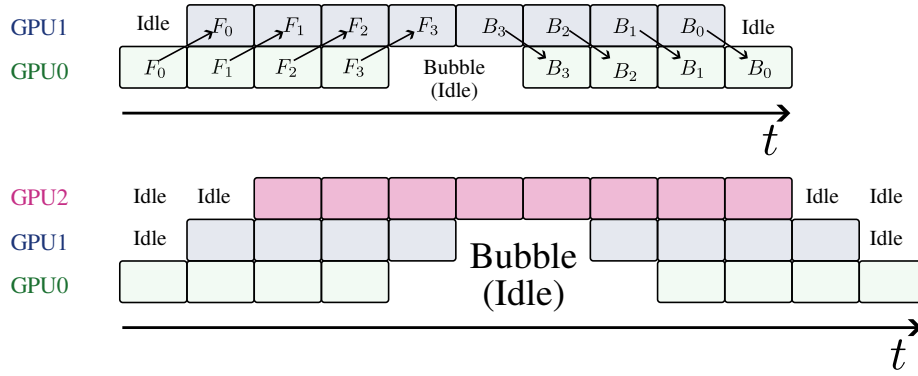


Figure 5.1: The Illustration of the pipeline parallel on two GPUs. As you can see the *bubble* (i.e., underutilization) tends to grow as we increase the number of GPUs

The input minibatch is then split into smaller micro-batches (smaller pieces of data), which flow sequentially through these sub-networks. In other words, the micro-batch is the basic unit of the input to the pipeline parallelism.

- While Stage 1 is processing the next micro-batch, Stage 2 can concurrently work on the intermediate outputs from Stage 1's previous micro-batch.

Example: Imagine a 2-stage pipeline parallel setup (for simplicity):

- GPU 0: Holds Layers 1–3
- GPU 1: Holds Layers 4–6

If you have a batch of data with 32 samples, you might split it into 4 micro-batches of size 8 each. Then, forward Pass can be processed as follows:

1. Micro-Batch 1
 - (a) Step A: GPU 0 processes layers 1–3 for micro-batch 1.
 - (b) Step B: Once GPU 0 is done with those layers, it sends the activations for micro-batch 1 over to GPU 1.
 - (c) Step C: GPU 1 then processes layers 4–6 for micro-batch 1.
2. Micro-Batch 2
 - (a) As soon as GPU 0 finishes Step A for micro-batch 1 and passes the data to GPU 1, GPU 0 is free to start micro-batch 2 (layers 1–3).
 - (b) Meanwhile, GPU 1 is busy processing micro-batch 1 (layers 4–6).
 - (c) Once GPU 0 finishes its part for micro-batch 2, it sends those activations to GPU 1—which will be ready to handle them as soon as it's done with micro-batch 1.
3. Micro-Batch 3 and 4
 - (a) This pattern continues in an overlapping fashion: while GPU 1 is busy with micro-batch 2, GPU 0 can start on micro-batch 3, and so on.

The key benefit is concurrency:

- While GPU 0 is processing micro-batch 2, GPU 1 can process micro-batch 1.
- This overlap leads to higher GPU utilization.

Backward pass is a bit more complex because:

- You need gradient signals to flow in the reverse order of the forward pipeline.
- Each stage waits until it receives the gradient from the next stage before it can compute its own local gradients and pass them back to the previous stage.

However, the overall concept is similar-multiple stages can run backprop (on different micro-batches) in parallel, thereby keeping all GPUs busy.

5.1.2 Pipeline Bubbles

When using pipeline parallelism, you often hear about *pipeline bubbles* (or underutilization). This refers to idle times on some GPUs before the assembly line is fully loaded or after it starts to wind down.

- Start-up Bubble: In the very beginning, GPU 1 must wait until GPU 0 finishes the first forward pass for micro-batch 1. GPU 1 sits idle during that initial delay.
- Wind-down Bubble: After the last micro-batch enters GPU 0, GPU 1 continues to process the pipeline while GPU 0 is idle.

The percentage of idle can be computed as follows:

$$\frac{1 - m}{m + n - 1},$$

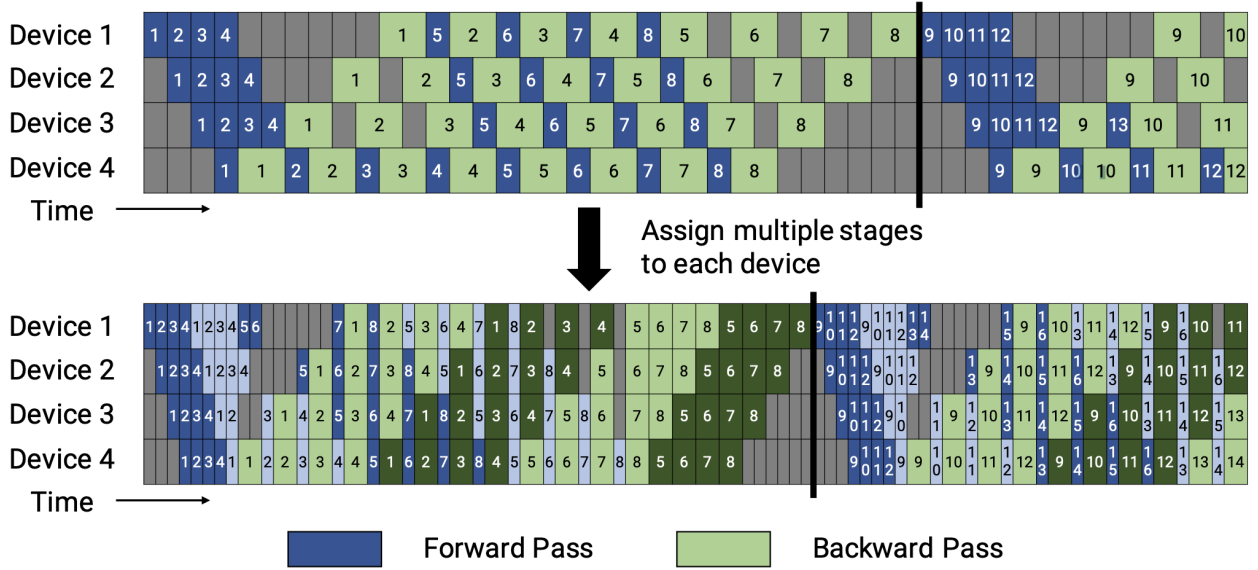
where m is the number of microbatches and n is the number of GPUs.

These bubbles can lead to less-than-ideal speedups, but you can mitigate them by using enough micro-batches to keep the pipeline busy most of the time.

5.1.3 Combining Pipeline Parallelism with Other Forms of Parallelism

In practice, pipeline parallelism is often combined with:

- Data Parallelism: You still replicate each stage across multiple GPUs to handle separate shards of data.
- Tensor Parallelism / Model Parallelism: Instead of giving entire layers to one GPU, you split the parameters or compute of a single layer across multiple GPUs (common in large language model setups, *e.g.*, Megatron-LM).
- Sharded Optimizer Approaches (*e.g.*, ZeRO, FSDP): Distribute optimizer states and gradients to reduce memory overhead.



5.2 1F1B

One of the issues is that the model parameters keep changing while processing the forward passes. This means at every time step, minibatches are going to be forwarded through different weights. Thus, it is necessary to keep different states of the model parameters. Thus, 1F1B increases the memory requirements while increasing the processing speed.

5.2.1 Non-interleaved Schedule

The non-interleaved schedule can be divided into two states. The first state is the startup state (or warm-up state). In the startup state, After completing the forward pass for the first minibatch, it performs the backward pass for the same minibatch, and then starts alternating between performing forward and backward passes for subsequent minibatches. As the backward pass starts propagating to earlier stages in the pipeline, every stage starts alternating between forward and backward pass for different minibatches. As shown in the above figure, in the steady state, every machine is busy either doing the forward pass or backward pass for a minibatch.

5.2.2 Interleaved Schedule

This schedule requires the number of microbatches to be an integer multiple of the stage of pipeline. In this schedule, each device can perform computation for multiple subsets of layers (called a model chunk) instead of a single contiguous set of layers. *i.e.*, Before device 1 had layer 1-4; device 2 had layer 5-8; and so on. But now device 1 has layer 1,2,9,10; device 2 has layer 3,4,11,12; and so on. With this scheme, each device in the pipeline is assigned multiple pipeline stages and each pipeline stage has less computation. This mode is both memory-efficient and time-efficient.

5.3 Zero Bubble

Chapter 6

Tensor Parallelism

6.1 Introduction

Let's go over an example:

- x is a row vector of shape $[1, d_{\text{in}}]$ (the input).
- W is a weight matrix of shape $[d_{\text{in}}, d_{\text{out}}]$.
- output is $[1, d_{\text{out}}]$.

We have two GPUs, GPU 0 and GPU 1. We want to split (shard) the weight matrix W across two GPUs. One common approach is column parallelism:

- GPU 0 holds columns $[0, 1]$
- GPU 1 holds columns $[2, 3]$

This means each GPU stores some columns of W . Let's denote:

$$W = [W_{\text{left}} \mid W_{\text{right}}]$$

where

- W_{left} is a 4×2 matrix on GPU 0,
- W_{right} is a 4×2 matrix on GPU 1.

In numeric form, suppose

$$W = \begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \\ 2 & 0 & 3 & 1 \\ -1 & 4 & 8 & 2 \end{bmatrix}.$$

Then, for column parallel:

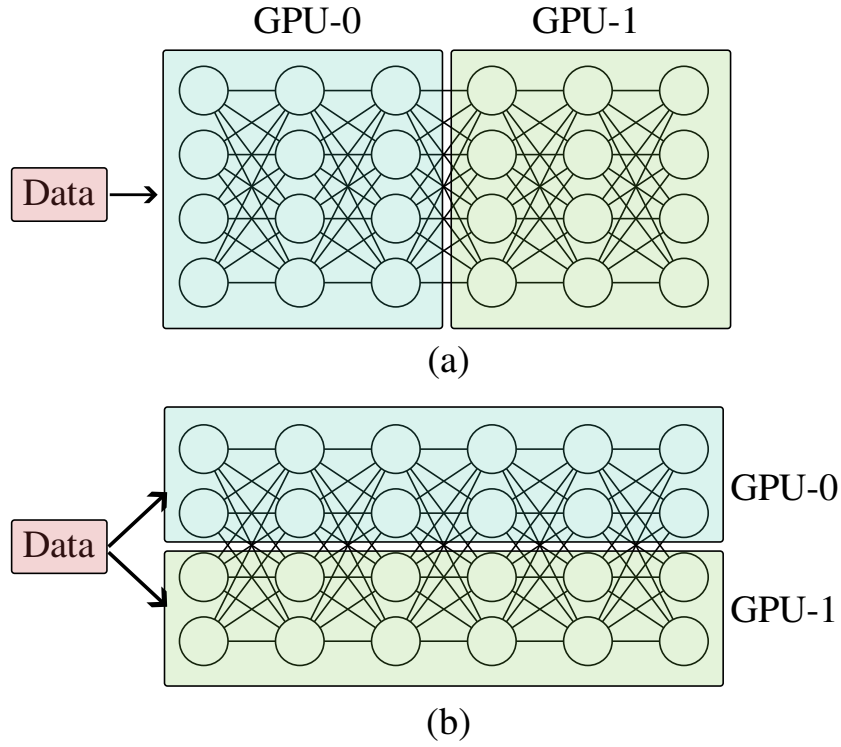


Figure 6.1: (a): Pipeline parallelism. (b) Tensor parallelism.

- GPU 0:

$$W_{\text{left}} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 2 & 0 \\ -1 & 4 \end{bmatrix}.$$

- GPU 1:

$$W_{\text{right}} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \\ 3 & 1 \\ 8 & 2 \end{bmatrix}.$$

Given the input

$$x = [1, 2, 0, 1].$$

We can treat x as a row vector $[1, 4]$. For column parallelism, each GPU needs the entire input x so it can multiply by its subset of columns:

- We copy the x to both GPU 0 and GPU 1.
 - This is typically a small overhead compared to storing large weight matrices.
- Then, compute the matrix multiplications for each matrix.
- Finally, concatenate the outputs.

$$\text{output} = [\text{partial}_0 \mid \text{partial}_1] = [6, 14, 27, 24].$$

- Some frameworks do a ring-all-gather, or they might place this final output on one GPU if needed, etc.

When we do backprop, we can update the model's parameters in the opposite direction.

In Megatron-LM, all Transformer layers, except normalization layer, are using row or column parallelism.

Tensor parallelism can be costly primarily due to the significant communication overhead involved when distributing large model layers across multiple GPUs, requiring frequent data exchange between devices which can become a bottleneck, especially when dealing with very large models and limited network bandwidth; this communication cost often outweighs the benefits of parallel computation, making it a major drawback of tensor parallelism.

Chapter 7

N -Dim Parallelism

Chapter 8

DualPipe

8.1 Introduction

8.1.1 All-to-All vs Point-to-Point

When orchestrating multiple GPUs, we need them to communicate with each other to share information like gradients and model parameters. There are two main types of communication patterns:

1. All-to-all communication.
2. Point-to-point communication.

All-to-all communication involves every GPU in the system simultaneously exchanging data with all other GPUs. The canonical analogy is a group chat where everyone needs to share their updates with everyone else. All-to-all communication is expensive and involves a ton of communication overhead. There are several clever algorithms like ring-AllReduce that can reduce this overhead, but it's still often a bottleneck.

Point-to-point communication, on the other hand, is a communication between just two GPUs (the analogy here is a private conversation). One GPU sends data directly to another specific GPU without involving the rest of the system. This is much more efficient in terms of network bandwidth and latency. In practice, point-to-point communication is strongly preferred when possible because it's significantly cheaper in terms of computational resources.

8.2 DualPipe

Finer-Grained stages: divide each chunk into 4 components:

- Attention,
- All-to-all dispatch(Handles communication between devices),
- MLP(Multi-Layer Perceptron),



Figure 8.1: An illustration of dualpipe.

- All-to-all combine(merge output across devices).

For a backward chunk, the attention and MLP split further into two parts: backward for input(B) and backward for weights(W) like Zero Bubble.

Bidirectional pipeline scheduling which feeds micro-batches from both ends of the pipeline simultaneously and a significant portion of communications can be fully overlapped (See the 2 black arrows in following diagram). In order to support bidirectional pipeline scheduling, DualPipe requires keeping two copies of the model parameters. If we have 8 devices with a 8 layers model, in the Zero Bubble Schedule, each device has a corresponding layer. But in the DualPipe Schedule, in order to handle bidirectional pipeline, the device 0 should have model's layer0 and layer7, and the device 7 should have model's layer7 and layer0.

Part III

Transformers

8.3 Flash Attention

Chapter 9

Tokenization

Part IV

Compression

Chapter 10

Model Compression

10.1 Introduction

haha

Bibliography

- [1] C. Brousseau and M. Sharp. *LLMs in Production: From Language Models to Successful Products*. Manning, 2025.
- [2] Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Zhengliang Liu, Zihao Wu, Peng Shu, Jie Tian, Tianze Yang, Shaochen Xu, Yanjun Lyu, Parker Blenk, Jacob Pence, Jason Rupram, Eliza Banu, Ninghao Liu, Linbing Wang, Wenzhan Song, Xiaoming Zhai, Kenan Song, Dajiang Zhu, Beiwen Li, Xianqiao Wang, and Tianming Liu. Large language models for manufacturing, 2024.