AI Traffic Amount Analysis

Lily Lyu (Huawei)

September Interim 2024

Purpose

In AICN study item report, one major comment is on traffic amount contents.

It is suggested to explain how we get the numbers.

<<Editor notes: the number in the table is highly dependent on the model architecture, parallel strategies: M=96, h=12288, L=96, b=1536, s=2048, T=8, P=8, D=16>>

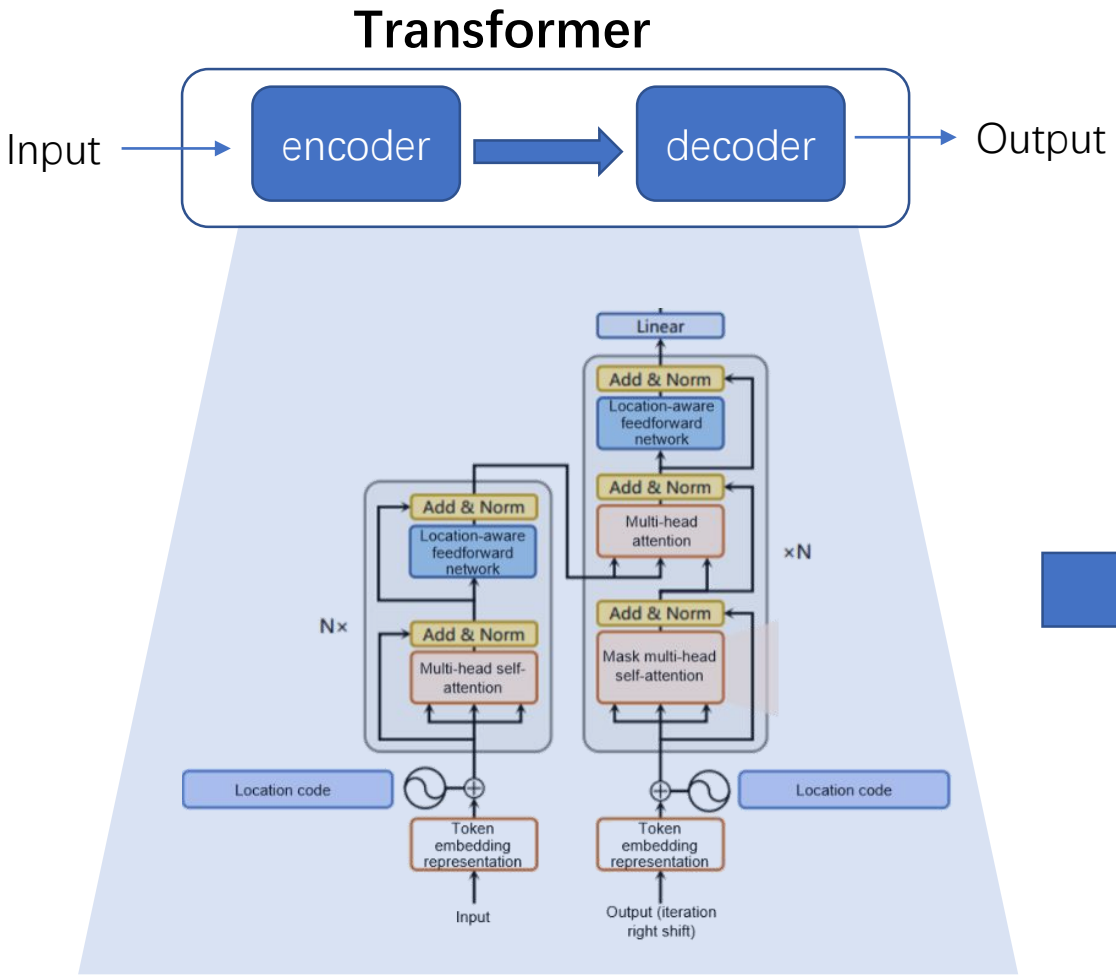
Taking GPT-3 175B model as an example, the amount of traffic generated by various types of communication in each iteration is listed in the following table. As shown by the table, TP mainly involves Allreduce operations. During each iteration of training, Allreduce operation are executed multiple times, depending on factors such as batch size, the number of neural network layers, pipeline lanes and parallel data streams. The amount of data to be communicated by the accelerator each time is determined by tensor lanes, batch size, hidden layer dimensions etc. Therefore, the total traffic amount communicated in each iteration for TP is the product of the number of communications and the data amount exchanged during each communication. That is hundreds of gigabytes.

	Typical operation	Total traffic amount
Tensor Parallelism (TP)	Allreduce	100s GB
Pipeline Parallelism (PP)	Send/receive	100s MB~10s GB
Data Parallelism (DP)	Allreduce	GB
Expert Parallelism (EP)	Alltoall	10s GB

https://mentor.ieee.org/802.1/dcn/24/1-24-0028-04-ICne-aicn-report-draft.pdf

This contribution intents to introduce the method for calculating communication traffic in typical types of parallelism.

Transformer Architecture Enables GenAI



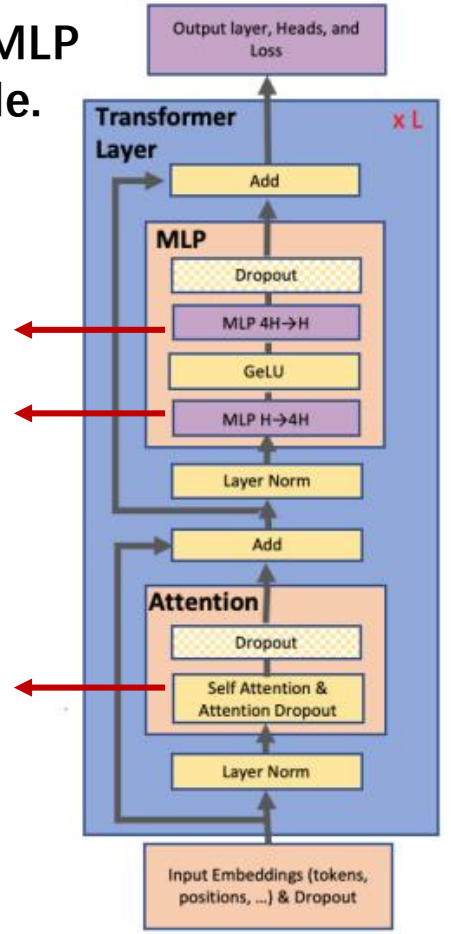
Structure of Transformer-based encoder and decoder

Each transformer layer has a MLP module and Attention module.

The number of MLP parameters is about **4h*h**

The number of MLP parameters is about **h*4h**

The number of Self-Attention parameters is about **4*h*h**, covering weight matrices Wq, Wk, Wv, and Wo

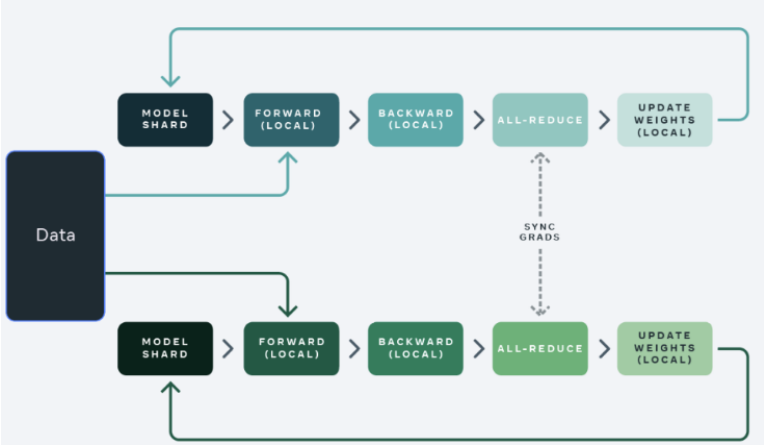


Source: Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism

Parallelism Policies Are Employed

Data Parallelism (DP)

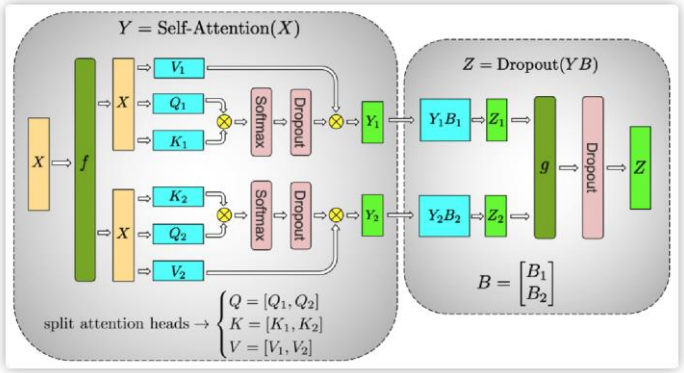
AllReduce for gradients sync

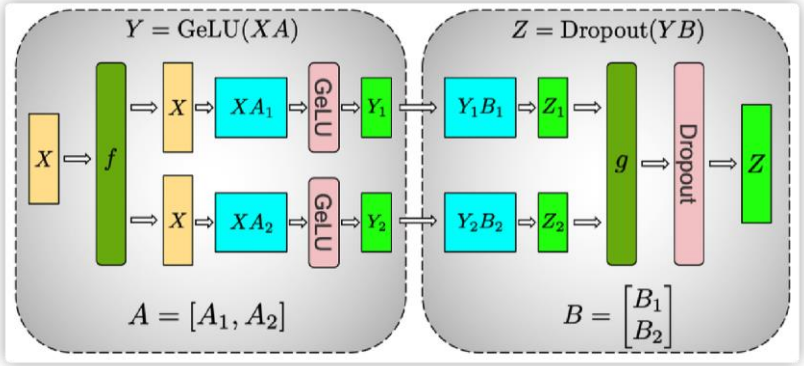


Tensor Parallelism (TP)

2 AllReduce for Attention and
2 AllReduce for MLP

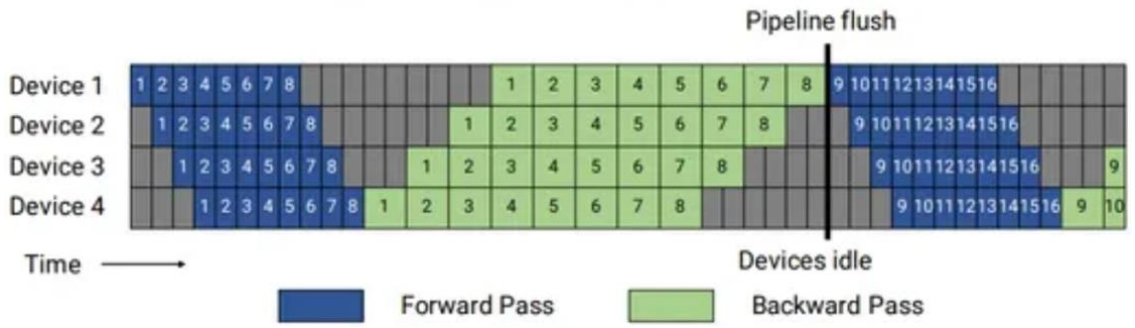
f: backward
g: forward





Pipeline Parallelism (PP)

Send/receive p times (p is the number of pipelines)



Example of Traffic Amount Calculation

GPT-3 example:

- L=96(layer number), h=12288(hidden dimension), b=1536(global batch size), s=2048(sequence length)
- T=8, P=8, D=16 (totally 8*8*16=1024 GPUs)
- AllReduce = reduce scatter + all gather, introduces 2 times traffic amount
- 2 bytes for each parameter

	Collective communication		GPU Traffic amount/time	Times/iteration	GPU traffic amount/iteration
DP	AllReduce	MLP	4h*h*2/T/D * 2(D-1) * 2byte = 540MB	L/P=12	12*(540+270) = **9.49GB**
		Attention	4h*h/T/D *2(D-1) * 2byte = 270MB	L/P=12	
PP	Send/Receive	Transformer	b/D * s * h * 2byte = 4.5GB	2	4.5*2 = **9GB**
TP	AllReduce	MLP	b/D*s*h/T * 2(T-1) * 2byte = 7.875GB	2 * L/P =24	7.875*24*2= **378GB**
		Attention	b/D*s*h/T * 2(T-1) * 2byte = 7.875GB	2 * L/P =24	

Note:

- Pipeline optimization is not used
- Input layer and output layer is not included in the calculation

Thanks!