A NOTATIONS

Throughout the discussion of modeling the training cost, we introduce the notations in Table 1.

Table 1: Table of notations.

Symbol	Description
d	GPU device.
$\mathbf{D}$	Set of N GPU devices.
$\mathbf{C}$	Tensor core of $\mathbf{D}$.
$\mathbf{M}$	Memory limit of $\mathbf{D}$.
m_d	GPU memory bandwidth of device d.
c_d	Tensor core computation power of device d.
$\mathbf{A}$	Communication matrix between devices describing the latency.
$\mathbf{B}$	Communication matrix between devices describing the bandwidth.
$\alpha_{d,d'}$	Latency between devices d and d'.
$\beta_{d,d'}$	Bandwidth between devices d and d'.
L	Total number of layers in the model.
H	Size of the hidden dimension in a transformer block.
B_{type}	Byte size for computational precision.
S	Sequence length in a transformer block.
B_{mb}	Micro-batch size.
n_{mb}	Number of micro-batches.
D_{pp}^{i}	Pipeline parallel degree for the i-th pipeline.
D_{dp}	Data parallel degree.
$\mathbf{d}_{i,j}$	Set of GPUs serves the j-th stage in the i-th pipeline.
$\lvert\mathbf{d}_{i,j}\rvert$	Tensor model parallel degree of the j-th stage in the i-th pipeline.
$\mathbf{D}_{\text{dp}}^{k}$	Set of GPUs serves the k-th transformer layer in data parallelism.
$l_{i,j}$	Number of layers on the i-th pipeline j-th stage.
$s_{i,j}$	Starting index of transformer layers on the i-th pipeline j-th stage.
σ	Assignment of devices in $\mathbf{D}$.

B COST ESTIMATION

In this section, we model the Comm-Cost, Comp-Cost, and Mem-Cumsum step by step. First we model cost for each transformer layer, and then model the end-to-end cost for the model as follows:

Modeling Cost Layer-wisely

- **Tensor model parallel communication cost**. Suppose activation recompute is enabled. A transformer layer is running over a set of GPU $\mathbf{d}_{i,j}$, the tensor model parallel communication cost for a micro-batch can be estimated by:

$$\text{Comm-TP-Layer}\left(\mathbf{d}_{i,j}\right) = 12 \cdot \max_{d \in \mathbf{d}_{i,j}} \sum_{d' \in \mathbf{d_{i,j}} - \{d\}} \left(\alpha_{d,d'} + \frac{B_{mb} S H B_{\text{type}}}{\left|\mathbf{d}_{i,j}\right| \beta_{d,d'}}\right) \tag{1}$$

- **Data parallel communication cost**. The data parallel communication cost for a transformer layer can be estimated by:

$$\text{Comm-DP-Layer}\left(\mathbf{D}_{\text{dp}}^{k}\right) = 2 \cdot \max_{d \in \mathbf{D}_{\text{dp}}^{k}} \sum_{d' \in \mathbf{D}_{\text{dp}}^{k} - \{d\}} \left(\alpha_{d,d'} + \frac{12 H^2 B_{\text{type}}}{\left|\mathbf{D}_{\text{dp}}^{k}\right| \beta_{d,d'}}\right) \tag{2}$$

- **Pipeline parallel communication cost**. Notice that pipeline parallel communication only happens when two layers are on different stages, e.g., the j-th stage and j+1-th stage in the i-th pipeline. It can be treated as two steps: the j-th stage `send` to the $j+1$-th stage in the forward pass and the $j+1$-th stage `broadcast` the information to every GPU in this stage (Or in the backward pass, the j-th stage `recv` from the $j+1$-th stage and then `broadcast`). Define $\text{Comm-PP-Hop}\left(\mathbf{d}_{i,D_{pp}^{i}}, \mathbf{d}_{i,D_{pp}^{i}+1}\right) = 0$ for convenience. The communication cost for a micro-batch can be estimated by:

$$\begin{aligned}\text{Comm-PP-Hop}\left(\mathbf{d}_{i,j}, \mathbf{d}_{i,j+1}\right) = 2 \cdot \min_{d \in \mathbf{d}_{i,j}, d' \in \mathbf{d}_{i,j+1}} \Bigg(& \left(\alpha_{d,d'} + \frac{B_{mb} S H B_{\text{type}}}{\beta_{d,d'}}\right) \\ & + \sum_{d'' \in \mathbf{d}_{i,j+1} - d'} \left(\alpha_{d',d''} + \frac{B_{mb} S H B_{\text{type}}}{\left|\mathbf{d}_{i,j+1}\right| \beta_{d',d''}}\right)\Bigg)\end{aligned} \tag{3}$$

- **Computation cost for a micro-batch**. Assume tensor model parallelism is always running over the same type of device d for any layer, and activation recomputation is enabled. The computation cost can be estimated by:

$$\textsc{Comp-TP-Layer}\left(\mathbf{d}_{i,j}\right) = \frac{96B_{mb}\cdot SH^2\left(1+\frac{S}{6H}\right)}{c_d|\mathbf{d}_{i,j}|} \tag{4}$$

Modeling Cost for Each Parallel Strategy

- **Data parallelism cost**. Different pipeline stages synchronize gradient simultaneously. The Data parallelism cost is bounded by the slowest pipeline stage, which can be estimated as:

$$\textsc{Comm-DP} = \max_{i,j}\left[\sum_{k=s_{i,j}}^{l_{i,j}}\textsc{Comm-DP-Layer}\left(\mathbf{D}_{\text{dp}}^{k}\right)\right] \tag{5}$$

- **Pipeline and tensor model parallelism cost**. For i-th pipeline to execute, the cost consists of the computation and communication cost for each stage (indexed by j):

$$\begin{aligned}&\textsc{Stage}\left(\mathbf{d}_{i,j}\right) = \\ &\sum_{k=1}^{l_{i,j}}\left[\textsc{Comp-TP-Layer}\left(\mathbf{d}_{i,j}\right)+\textsc{Comm-TP-Layer}\left(\mathbf{d}_{i,j}\right)\right]\end{aligned} \tag{6}$$

Notice that the slowest stage bounds the pipeline parallel stage. Thus, we formulate the pipeline and tensor model parallelism cost as below:

$$\begin{aligned}&\textsc{Pipeline-Time}\left(i\right) = \\ &\sum_{j=1}^{D_{pp}^{i}}\left(\textsc{Stage}\left(\mathbf{d}_{i,j}\right)+\textsc{Comm-PP-Hop}\left(\mathbf{d}_{i,j},\mathbf{d}_{i,j+1}\right)\right) \\ &+\left(n_{mb}-1\right)\cdot\max_{j=2,\ldots,D_{pp}^{i}}\left(\textsc{Stage}\left(\mathbf{d}_{i,j}\right)\right. \\ &\left.+\textsc{Comm-PP-Hop}\left(\mathbf{d}_{i,j},\mathbf{d}_{i,j+1}\right)\right)\end{aligned} \tag{7}$$

Modeling End-to-end time: One iteration time is determined by the slowest pipeline and the data parallel cost, which can be estimated as follows:

$$\begin{aligned}&\textsc{Comm-Cost}\left(\sigma\right)+\textsc{Comp-Cost}\left(\sigma\right) \\ &\quad=\max_{i=1,\ldots,D_{dp}}\textsc{Pipeline-Time}\left(i\right)+\textsc{Comm-DP}\end{aligned} \tag{8}$$

Modeling Memory Cost: Suppose full activation recompute and naive data parallelism are applied. The memory cost of parameters and activations can be estimated as below:

$$\textsc{Mem-Cumsum}\left(\sigma\right) = \frac{48H^2B_{type}}{|\mathbf{d}_{i,j}|}+B_{mb}SHB_{type} \tag{9}$$