Tutorial. May 21, 2023 2-6 pm

Distributed Training of Deep Neural Networks

Abhinav Bhatele, Siddharth Singh
Department of Computer Science

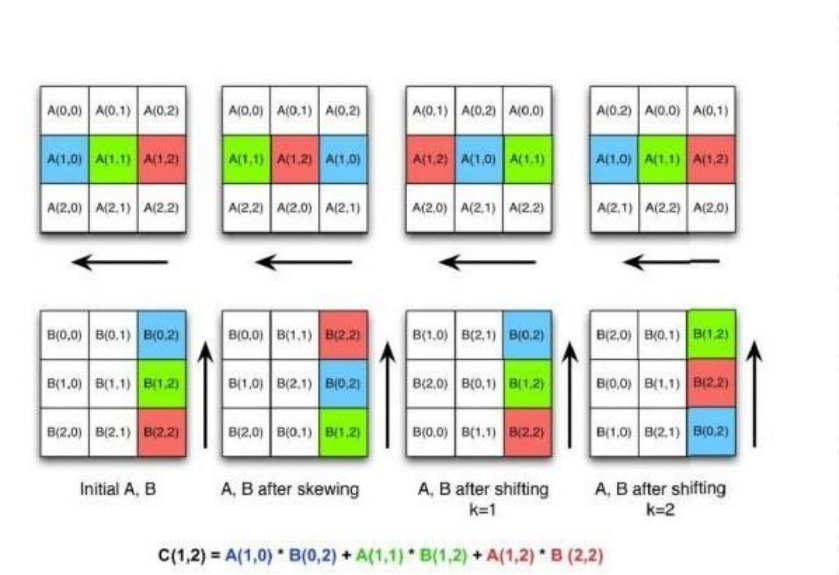
Limitations of data parallelism

- DDP – Supports models of limited size
- Deepspeed – Higher stages are inefficient

Intra-layer parallelism

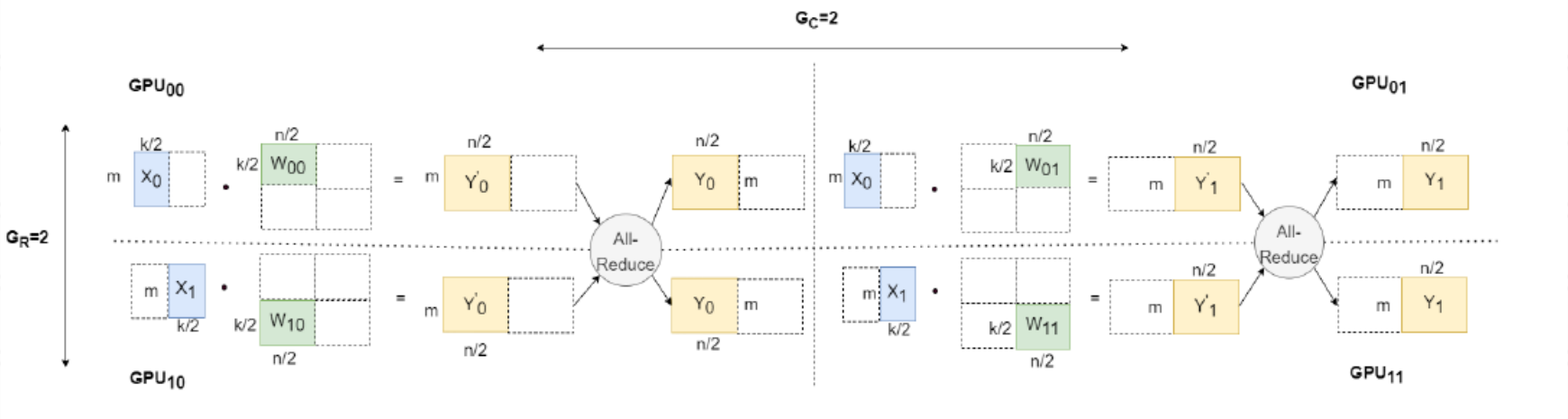
- Divide parameters and compute of every layer of a neural network on multiple GPUs.
- Two kinds of layers
    - ReLU and Layernorm – apply same function to each element of the input tensor
    - Fully Connected/Convolution – matrix multiplication operations that aren't easy to parallelize

Parallelizing a Matrix Multiplication



Cannon's Parallel Matrix Multiplication Algorithm

AxoNN's 2D Tensor Parallelism



Parallelizing a matrix multiplication (X.Y=Z) using AxoNN on 4 GPUs

Running the code

- Code – `train_axonn_intra_layer.py`

```
cd session_3_intra_layer_parallelism
sbatch --reservation=2023 run.sh
```