Tutorial. May 12, 2024 2-6 pm

Distributed Training of Deep Neural Networks

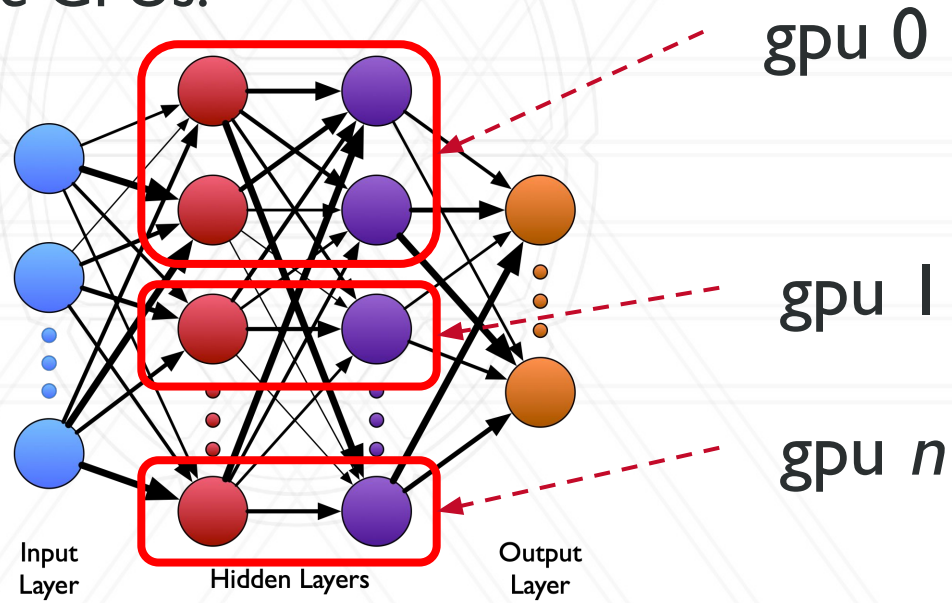
Abhinav Bhatele, Siddharth Singh, Daniel Nichols
Department of Computer Science

Limitations of data parallelism

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

Intra-layer parallelism

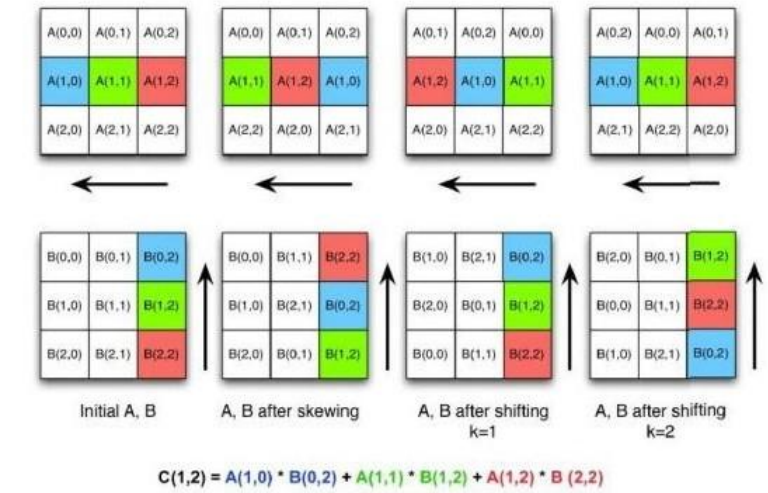
- Divide parameters and compute of every layer of a neural network on multiple GPUs.



Intra-layer parallelism

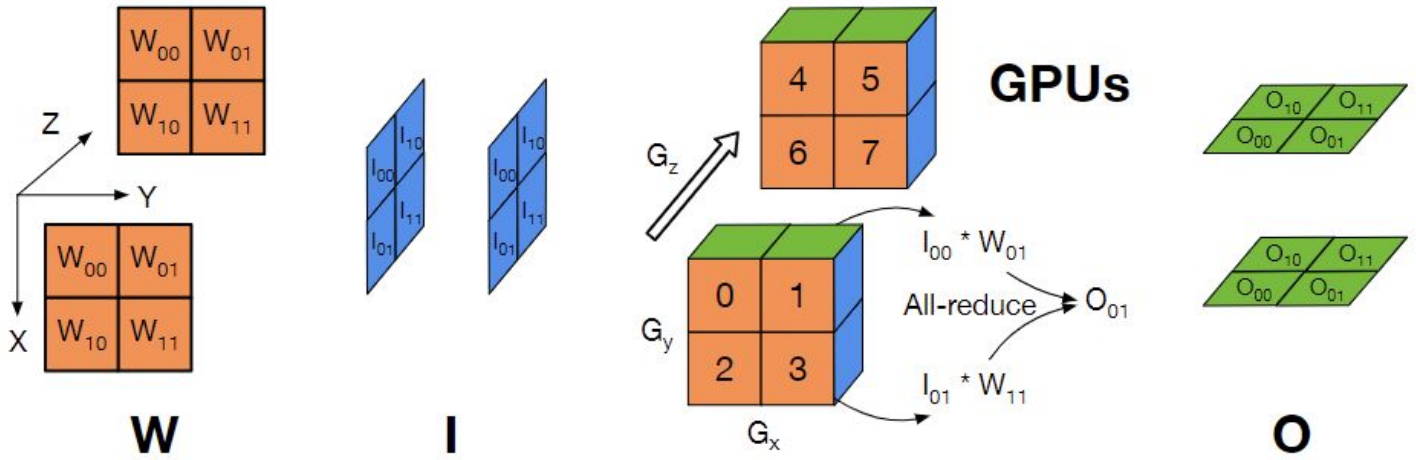
- Divide parameters and compute of every layer of a neural network on multiple GPUs.
- Two kinds of layers
    - ReLU and Layernorm – apply same function to each element of the input tensor
    - Fully Connected/Convolution – matrix multiplication operations that aren't easy to parallelize

Parallelizing a Matrix Multiplication



Cannon's Parallel Matrix Multiplication Algorithm

AxoNN's 3D Tensor Parallelism



Parallelizing a matrix multiplication (I.W=O) using AxoNN on 8 GPUs

Extremely easy to use!

```
from axonn.intra_layer import auto_parallelize
with auto_parallelize():
    net = FC_Net(args.num_layers, args.image_size**2, args.hidden_size, 10).cuda()
```

Zero code changes required in your model definition!

Running the code (Tensor/Intra-Layer)

- Code – `train.py`

```
cd session_3_intra_layer_parallelism
sbatch --reservation=isc2024 run.sh
```