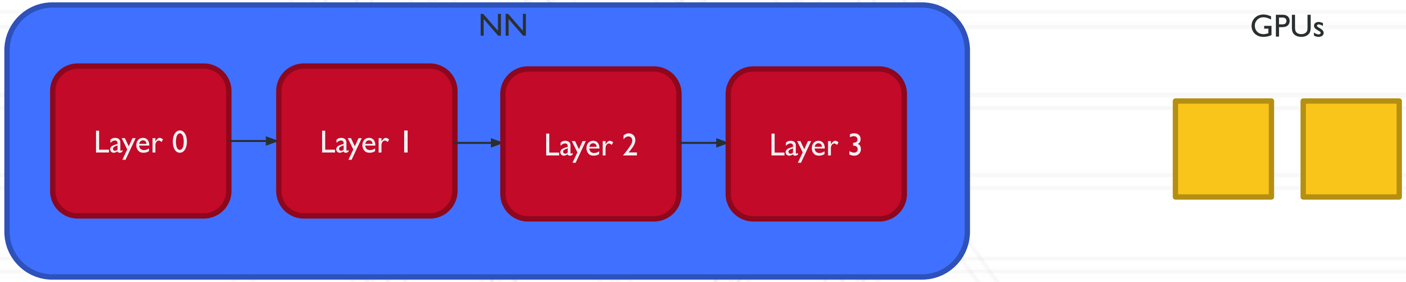
Tutorial. May 12, 2024 2-6 pm

Distributed Training of Deep Neural Networks

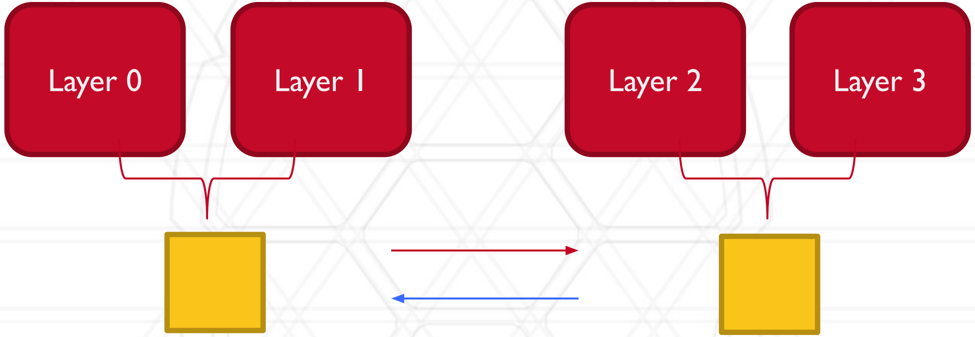
Abhinav Bhatele, Siddharth Singh, Daniel Nichols
Department of Computer Science

Inter-layer parallelism

- Map contiguous subsets of layers to GPUs.



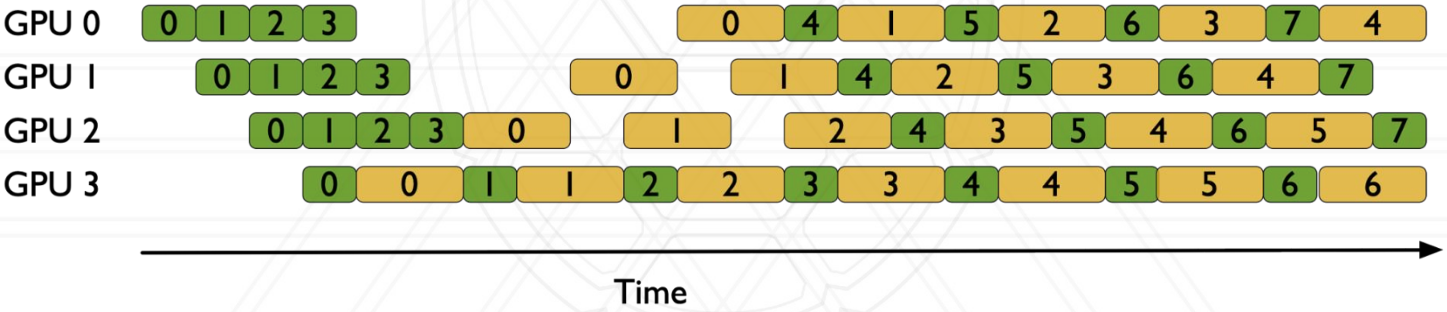
Inter-layer parallelism



Point-to-point communication of activations and their gradients

Inter-layer 'parallelism' ?

- Break batch intro micro-batches.
- Process micro-batches in a pipelined fashion.


GPU 0 0 1 2 3 0 4 1 5 2 6 3 7 4
GPU 1 0 1 2 3 0 1 4 2 5 3 6 4 7
GPU 2 0 1 2 3 0 1 2 4 3 5 4 6 5 7
GPU 3 0 0 1 1 2 2 3 3 4 4 5 5 6 6
Time

Running the code

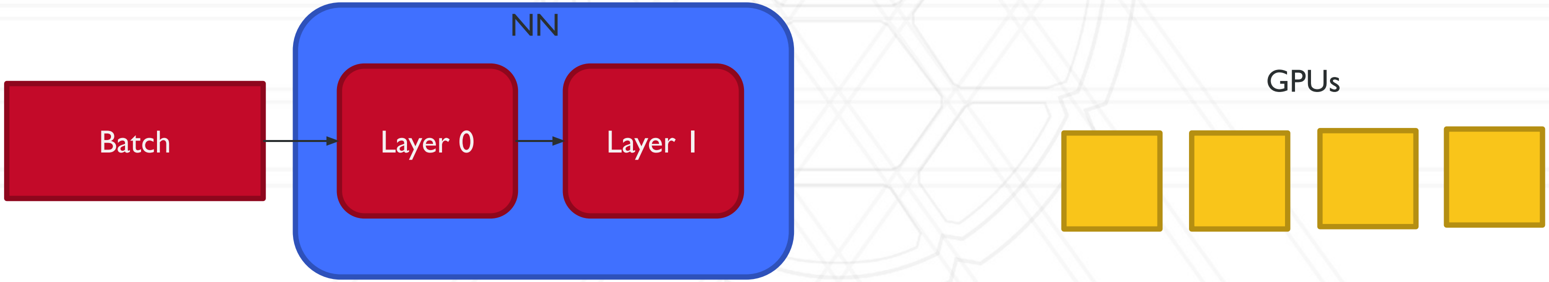
- Code – `train_axonn_inter_layer.py`

```
cd session_4_inter_layer_parallelism
sbatch --reservation=isc2024 run.sh
```

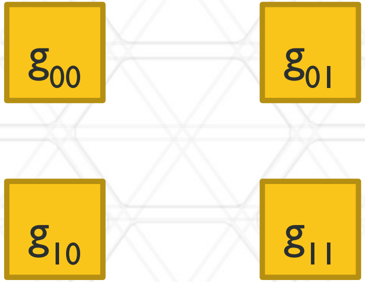
Hybrid parallelism

- AxoNN can combine inter/intra-layer parallelism with data parallelism too.

- Let us discuss inter-layer + data parallelism.

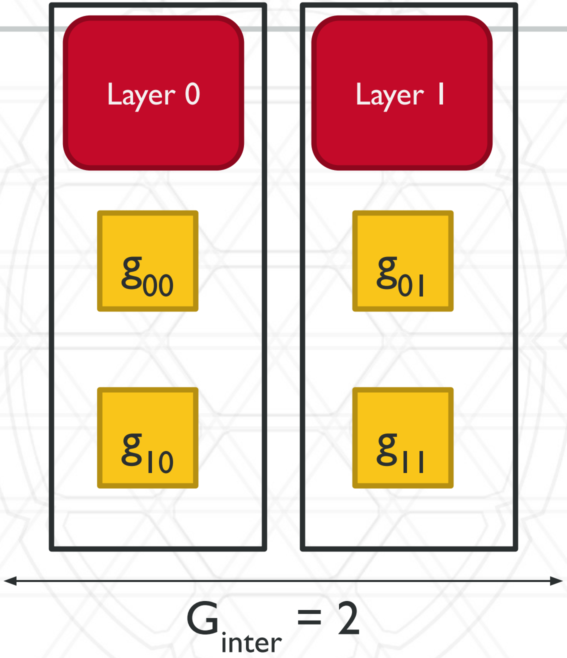


Inter-Layer+Data Parallelism



Organize GPUs in a 2D grid

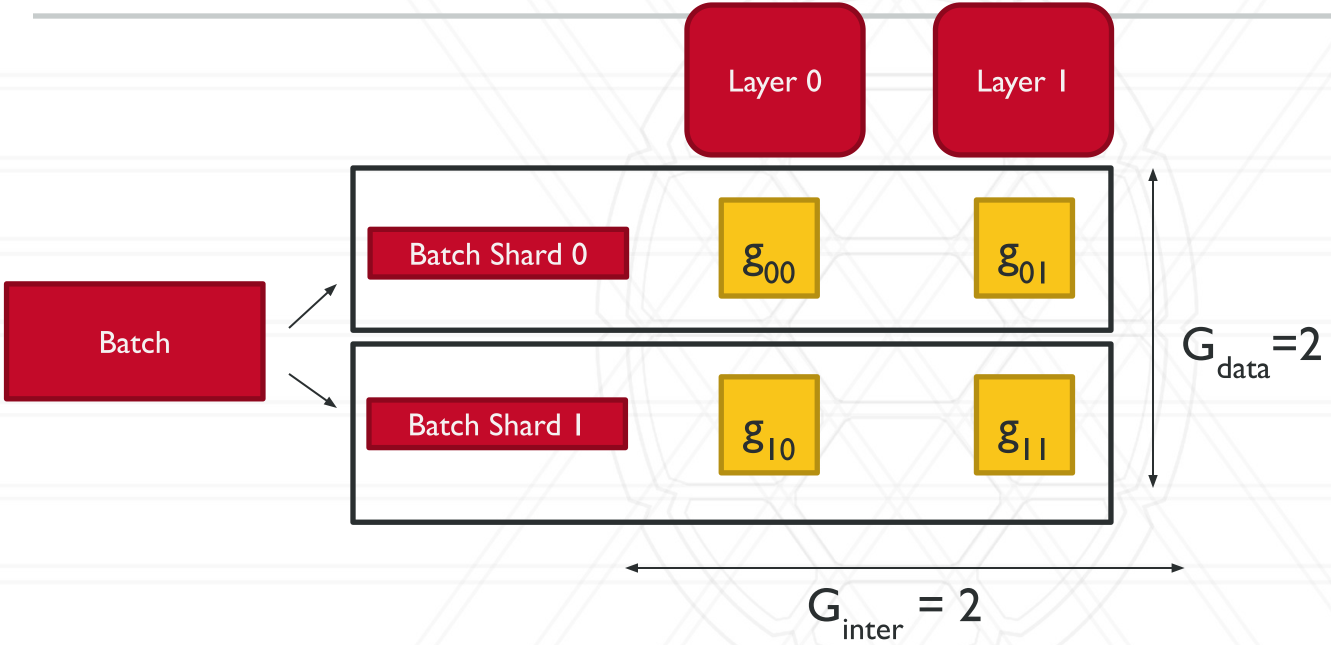
Inter-Layer+Data Parallelism



Inter-Layer Parallelism

Partition layers equally across columns

Inter-Layer+Data Parallelism



Data Parallelism

Partition batch equally across rows

Running the code

- Code – `train_axonn_inter_layer.py`

```
cd session_4_inter_layer_parallelism
HYBRID_PARR=true sbatch --reservation=isc2024 run.sh
```