IMAGINE TOMORROW
ISC High Performance

Tutorial. May 21, 2023 2-6 pm

Distributed Training of Deep Neural Networks

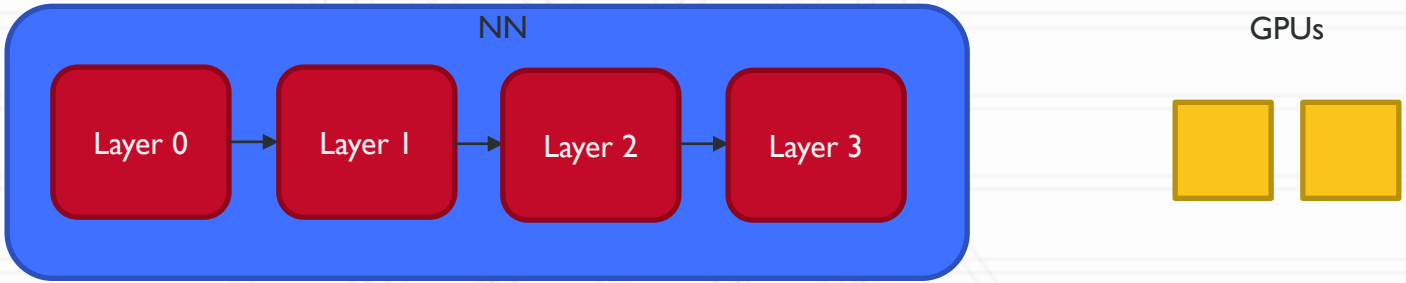
Abhinav Bhatele, Siddharth Singh
Department of Computer Science

PARALLEL SOFTWARE AND SYSTEMS GROUP

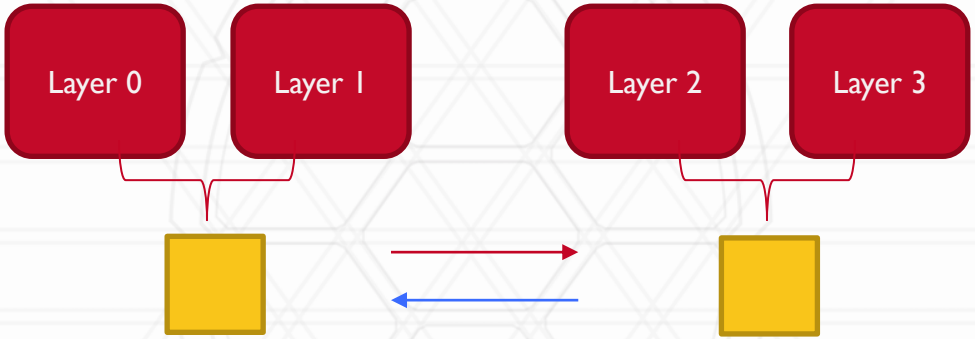
UNIVERSITY OF MARYLAND

Inter-layer parallelism

- Map contiguous subsets of layers to GPUs.



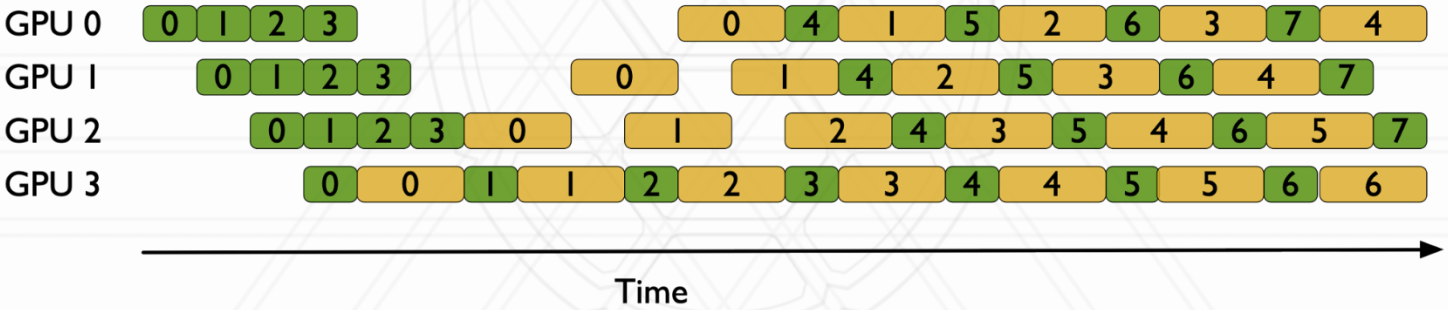
Inter-layer parallelism



Point-to-point communication of activations and their gradients

Inter-layer 'parallelism' ?

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



Running the code

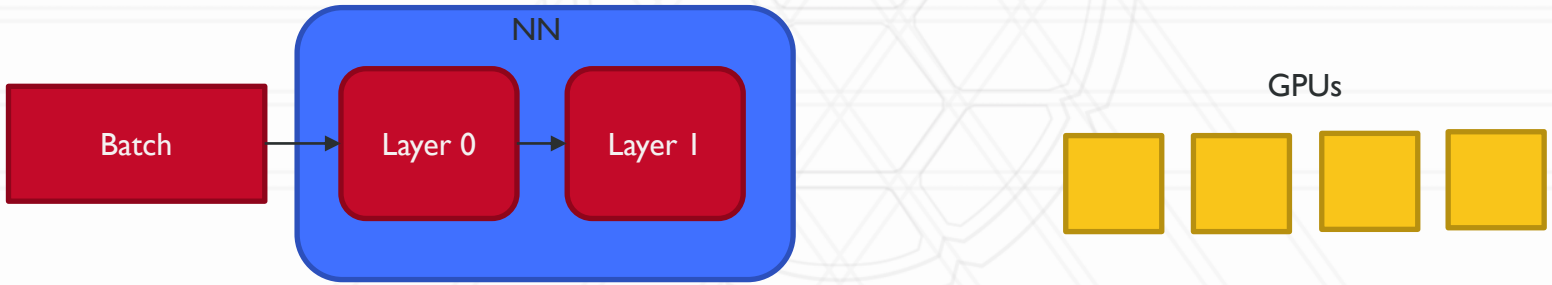
- Code – `train_axonn_inter_layer.py`

```
cd session_4_inter_layer_parallelism
sbatch run.sh
```

Hybrid parallelism

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

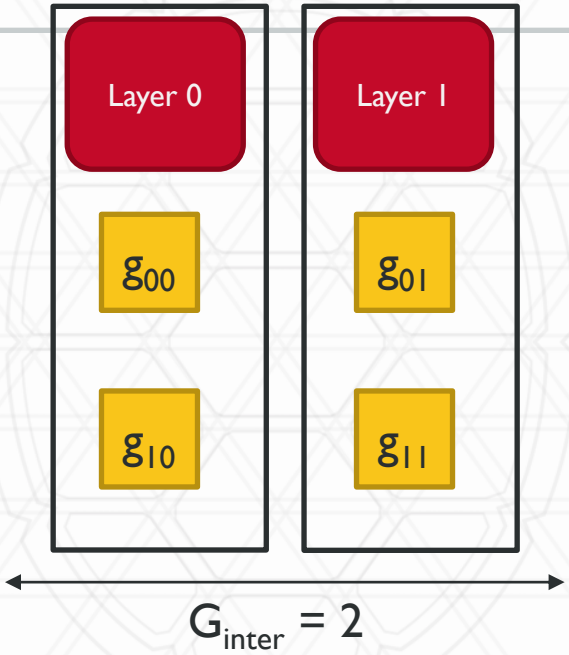
- Let us discuss inter-layer + data parallelism.



Inter-Layer+Data Parallelism

Organize GPUs in a 2D grid

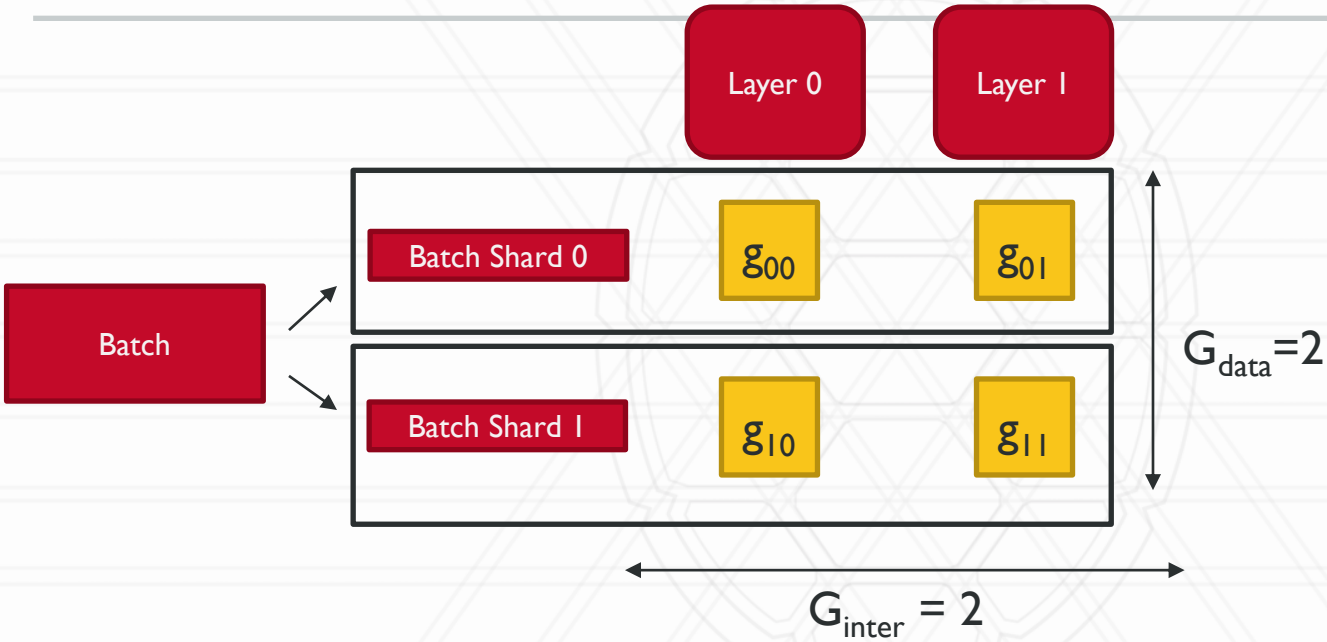
Inter-Layer+Data Parallelism



Inter-Layer Parallelism

Partition layers equally across columns

Inter-Layer+Data Parallelism



Partition batch equally across rows

Running the code

- Code – `train_axonn_inter_layer.py`

```
cd session_4_inter_layer_parallelism
HYBRID_PARR=true sbatch run.sh
```