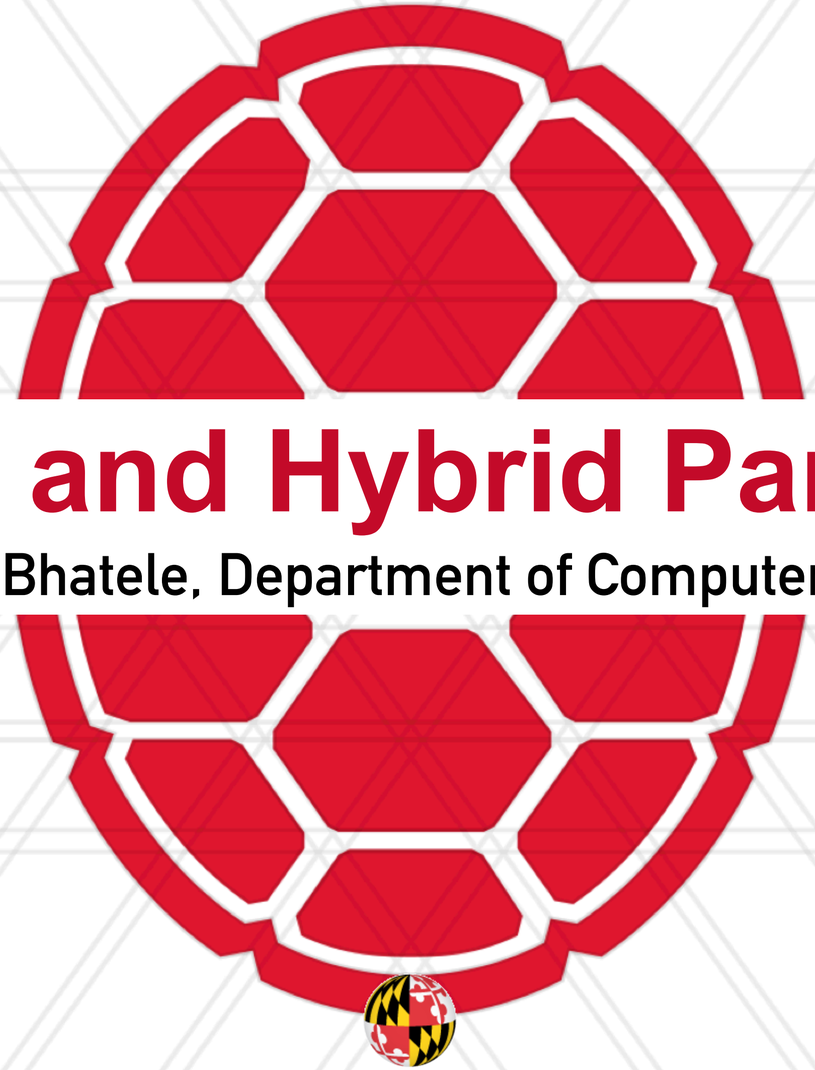
Systems for Machine Learning (CMSC828G)



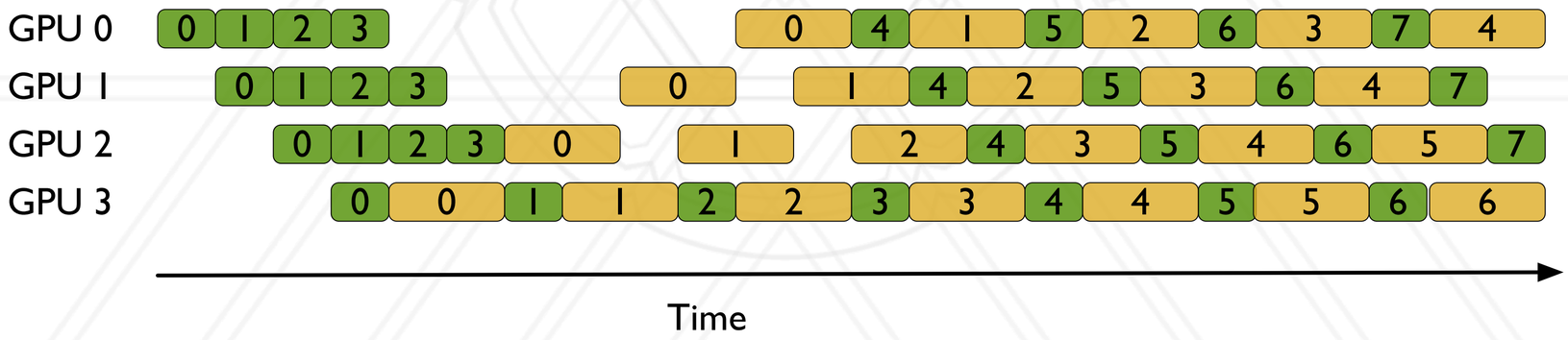
Pipeline and Hybrid Parallelism

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Inter-layer parallelism

- Assign entire layers to different processes/GPUs
  - Ideally map contiguous subsets of layers
- Point-to-point communication (activations and gradients) between processes/GPUs managing different layers
- Use a pipeline of mini-batches to enable concurrent execution

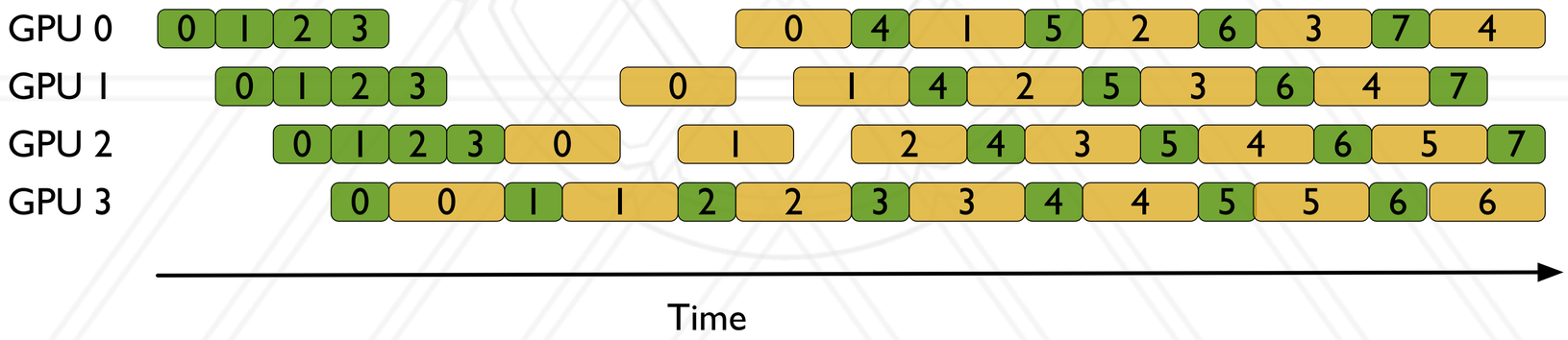


Inter-layer parallelism

- Assign entire layers to different processes/GPUs
  - Ideally map contiguous subsets of layers

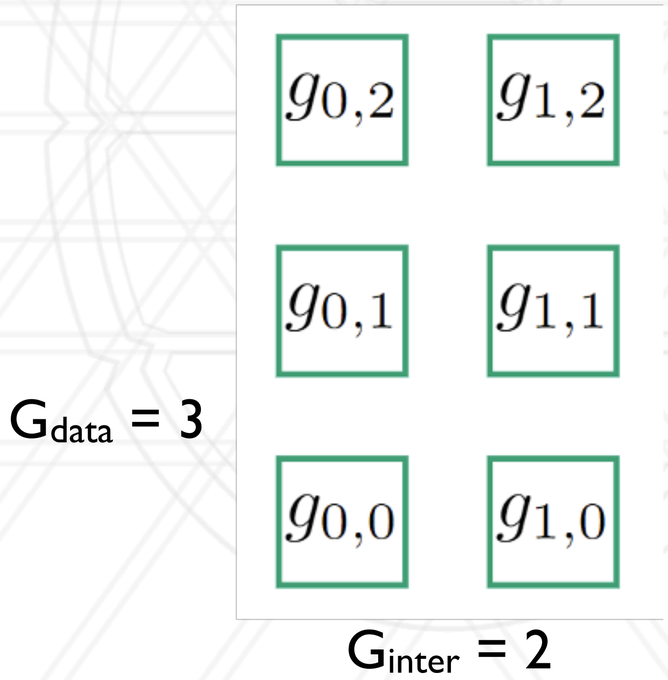
Pipeline parallelism

- Point-to-point communication (activations and gradients) between processes/GPUs managing different layers
- Use a pipeline of mini-batches to enable concurrent execution



Data + inter-layer parallelism

- 2D hybrid parallelism: data + inter-layer with pipelining

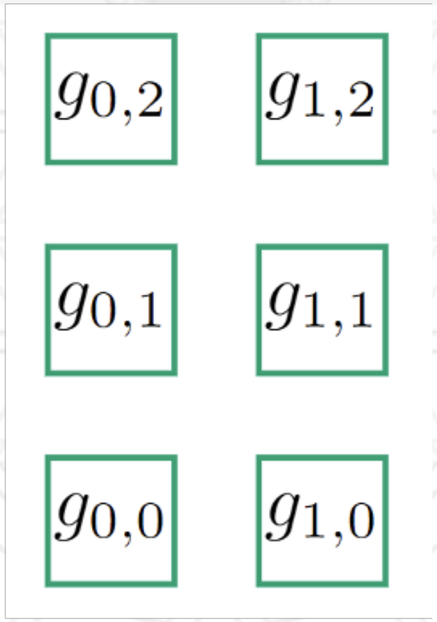


Data + inter-layer parallelism

- 2D hybrid parallelism: data + inter-layer with pipelining

Batch

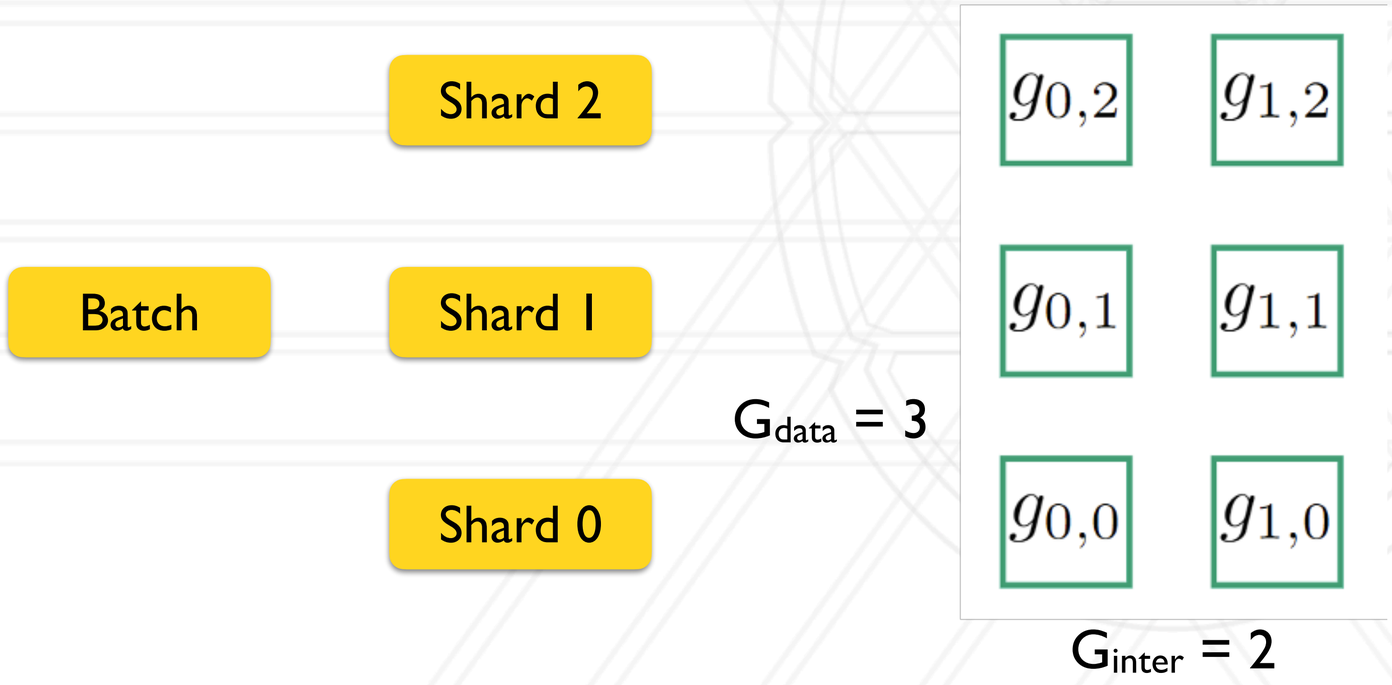
$G_{data} = 3$



$G_{inter} = 2$

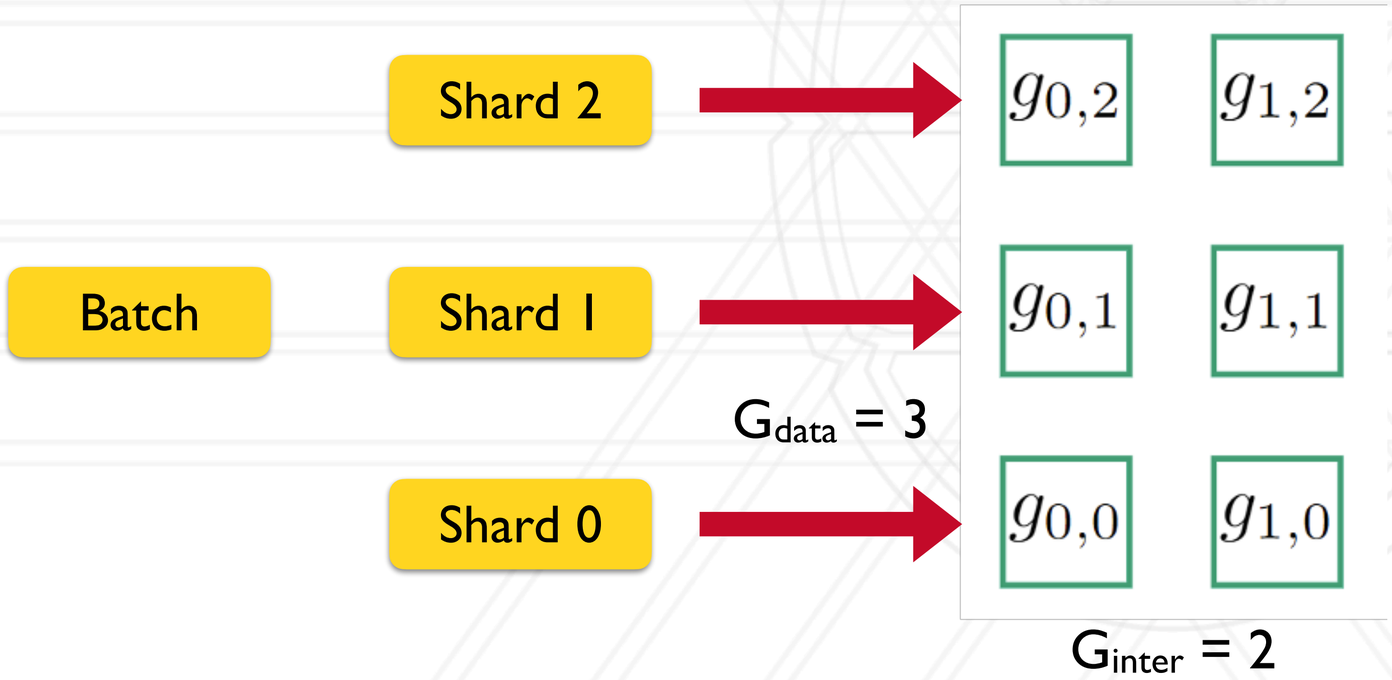
Data + inter-layer parallelism

- 2D hybrid parallelism: data + inter-layer with pipelining



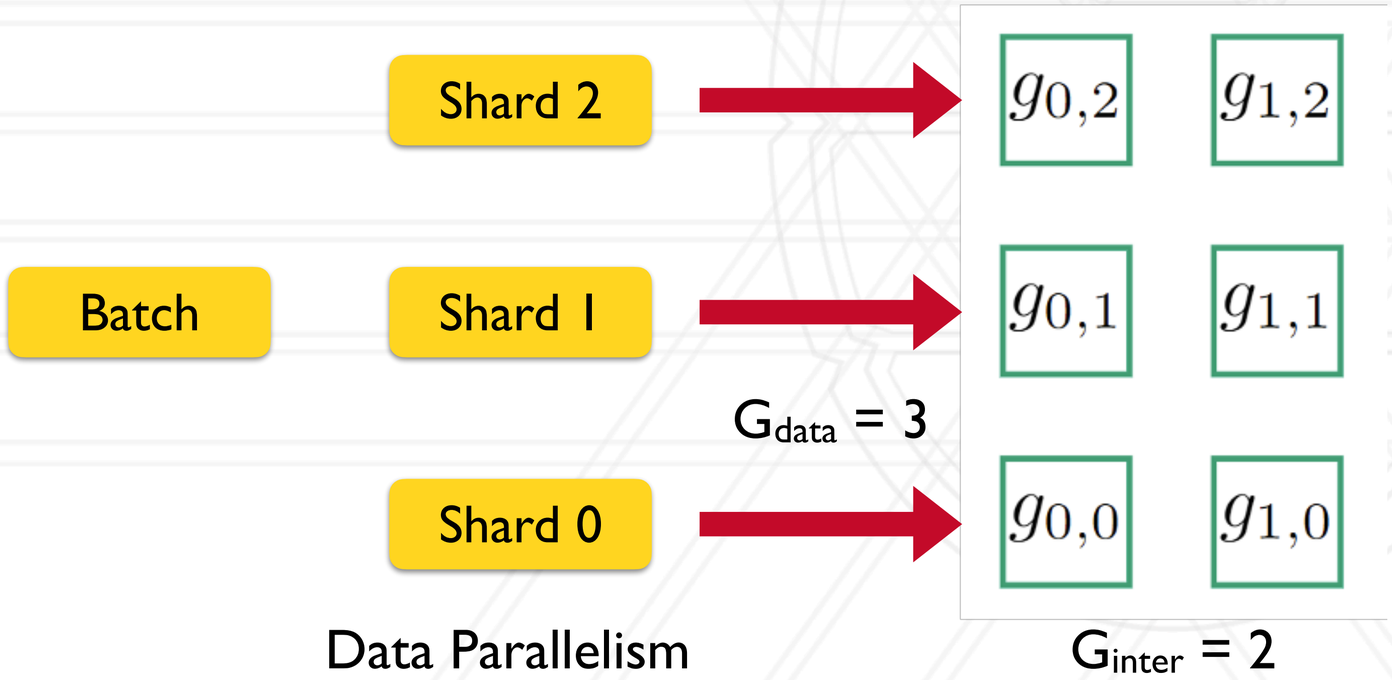
Data + inter-layer parallelism

- 2D hybrid parallelism: data + inter-layer with pipelining



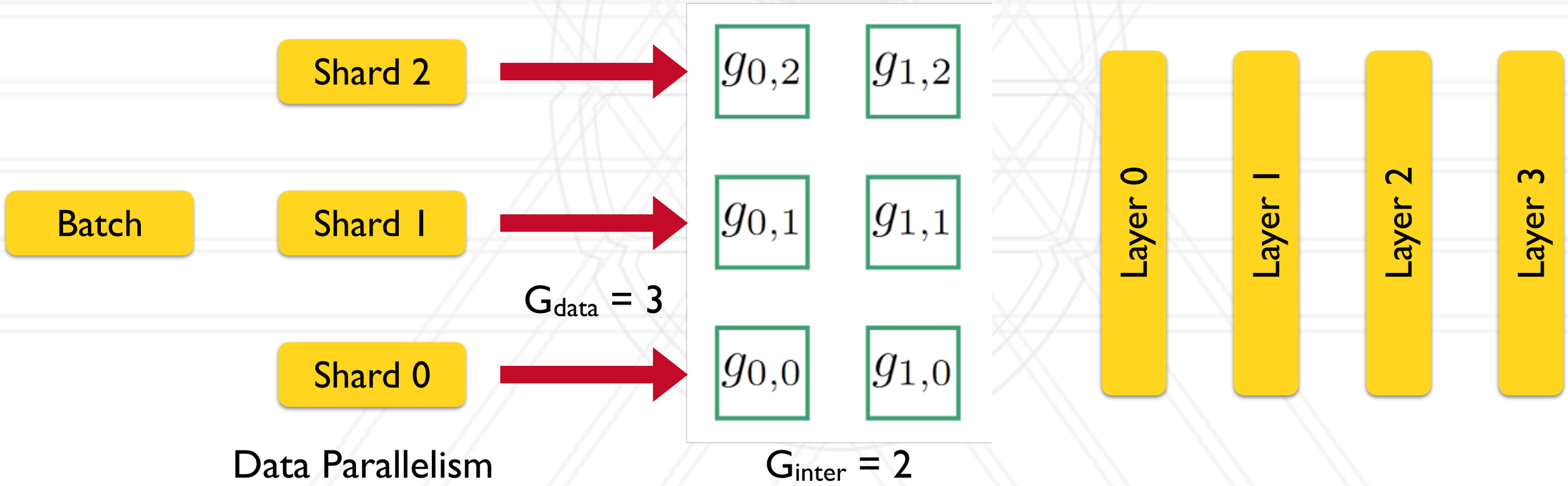
Data + inter-layer parallelism

- 2D hybrid parallelism: data + inter-layer with pipelining



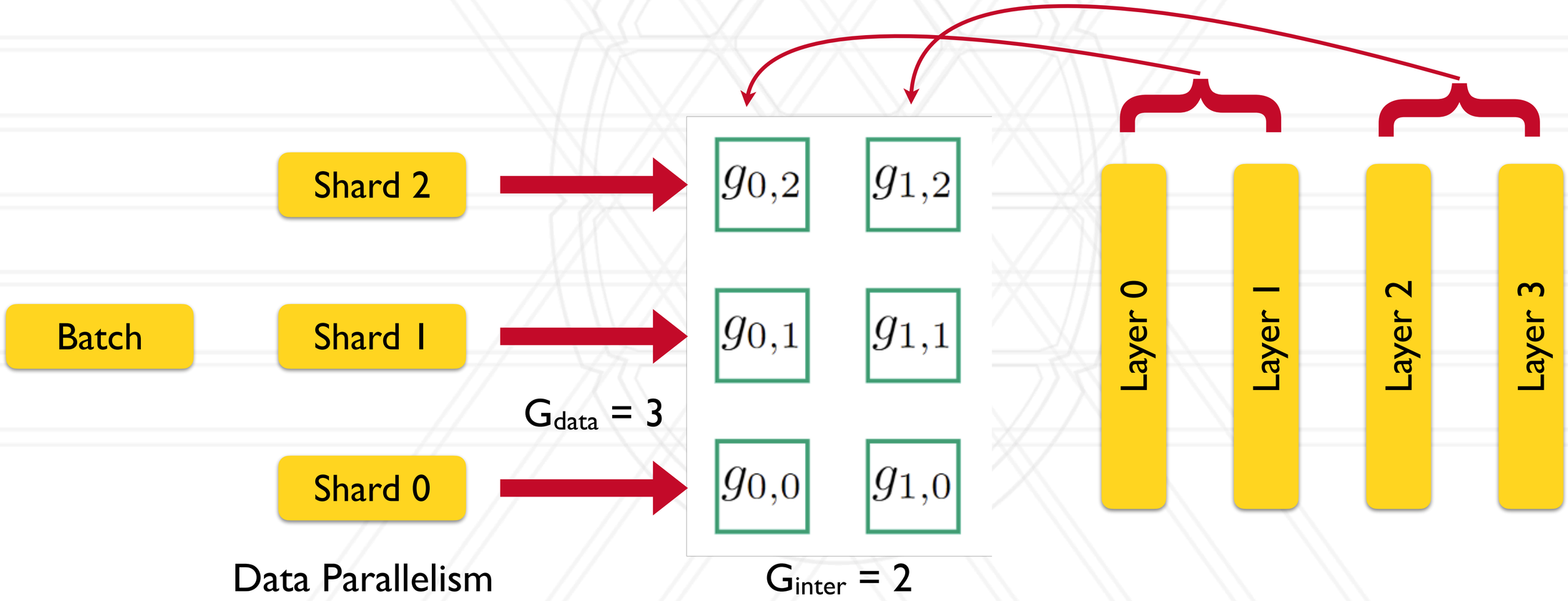
Data + inter-layer parallelism

- 2D hybrid parallelism: data + inter-layer with pipelining



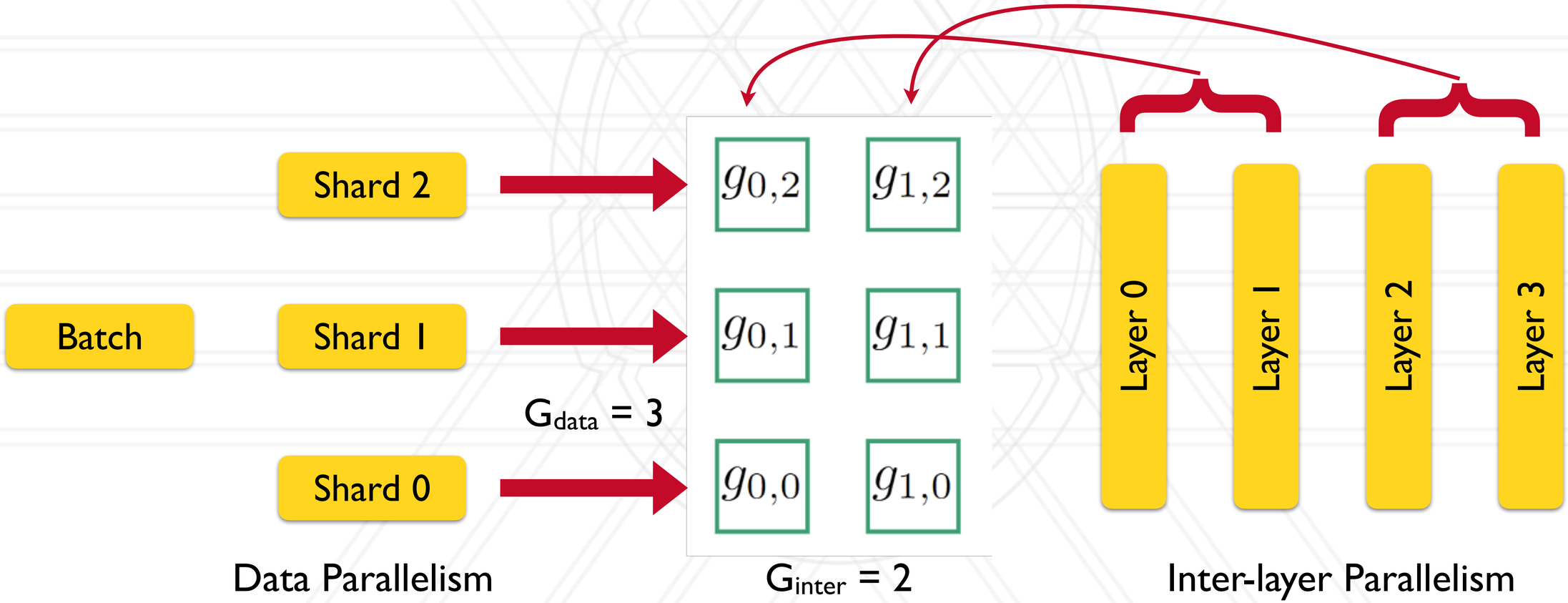
Data + inter-layer parallelism

- 2D hybrid parallelism: data + inter-layer with pipelining



Data + inter-layer parallelism

- 2D hybrid parallelism: data + inter-layer with pipelining



Hybrid parallelism

- Using two or more approaches together in the same parallel framework
- 3D parallelism: use all three
- Popular serial frameworks: pytorch, tensorflow
- Popular parallel frameworks: DDP, MeshTensorFlow, Megatron-LM, ZeRO, AxoNN

Hybrid data + X parallel approach

- Combines data parallelism with other forms of parallelism
- In general, you can combine all of data + tensor +pipeline (+ expert) parallelism

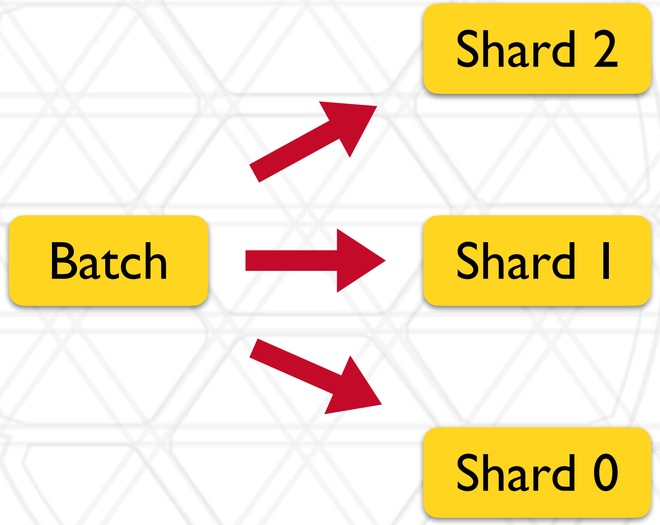
Hybrid data + X parallel approach

- Combines data parallelism with other forms of parallelism
- In general, you can combine all of data + tensor +pipeline (+ expert) parallelism

Batch

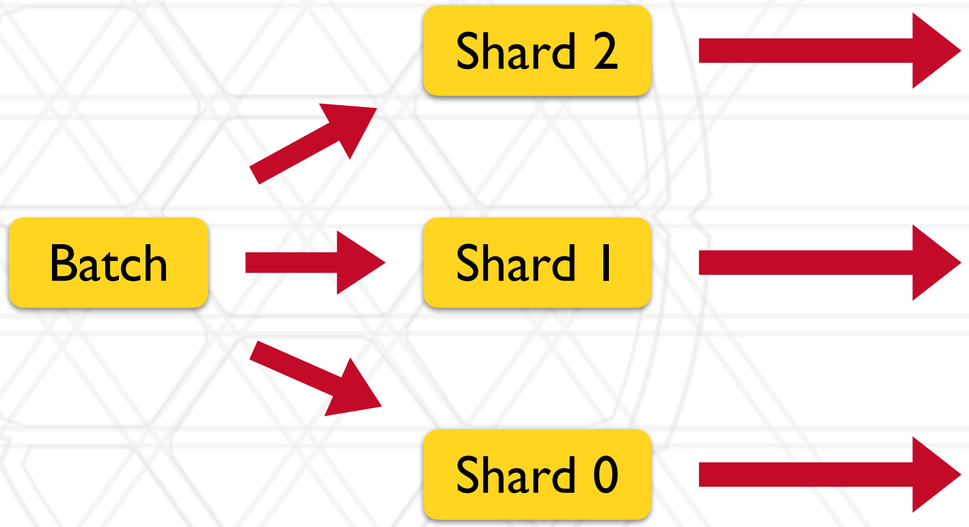
Hybrid data + X parallel approach

- Combines data parallelism with other forms of parallelism
- In general, you can combine all of data + tensor +pipeline (+ expert) parallelism



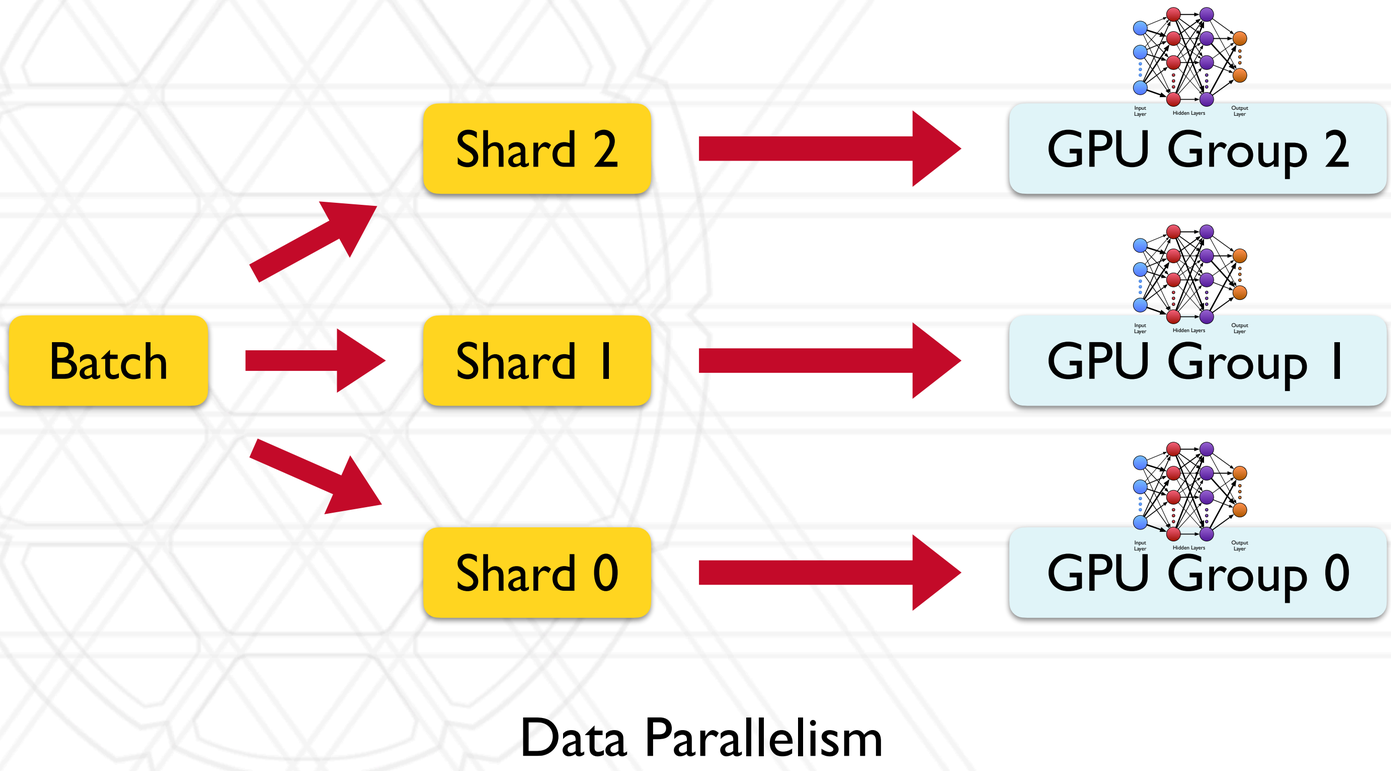
Hybrid data + X parallel approach

- Combines data parallelism with other forms of parallelism
- In general, you can combine all of data + tensor +pipeline (+ expert) parallelism



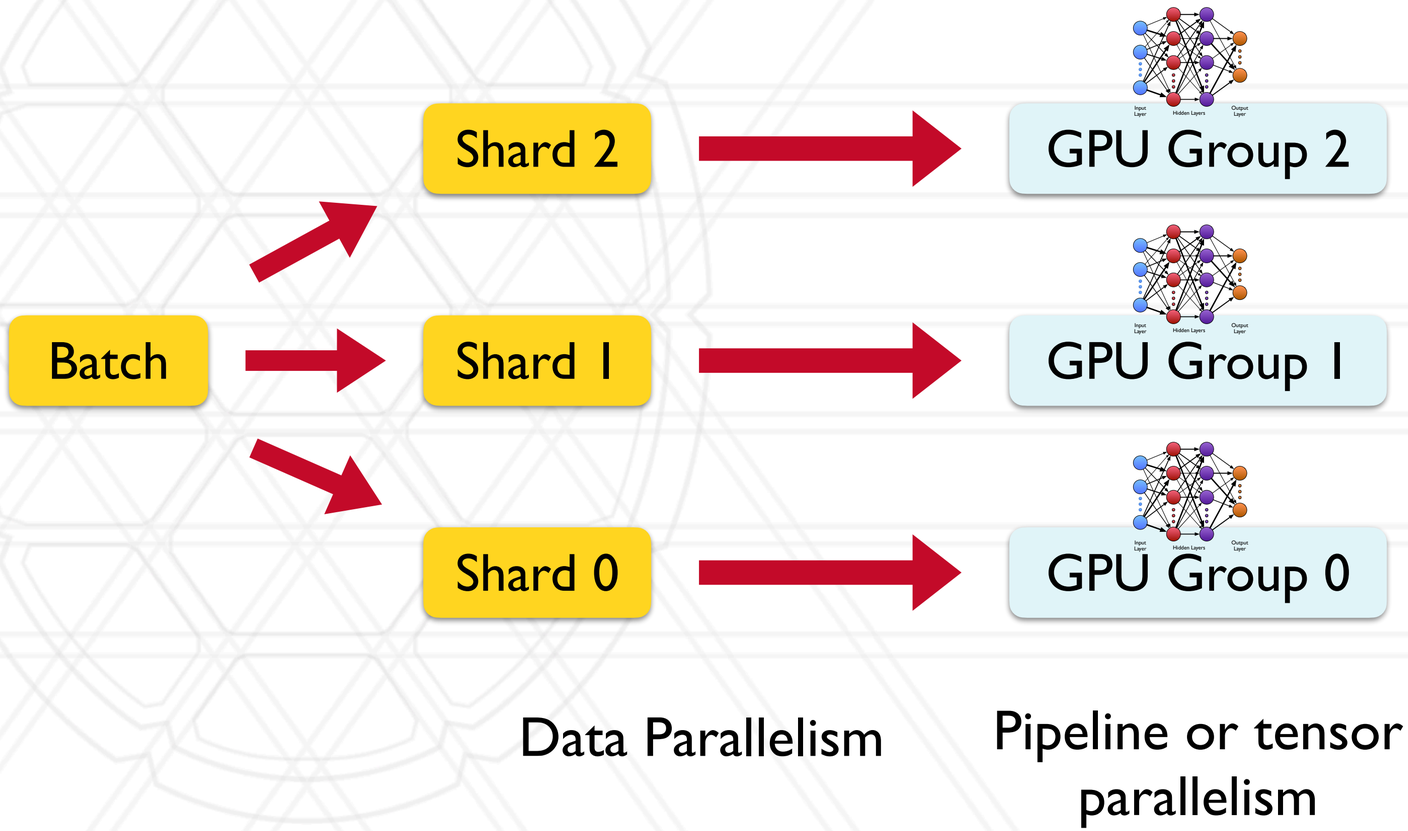
Hybrid data + X parallel approach

- Combines data parallelism with other forms of parallelism
- In general, you can combine all of data + tensor +pipeline (+ expert) parallelism



Hybrid data + X parallel approach

- Combines data parallelism with other forms of parallelism
- In general, you can combine all of data + tensor +pipeline (+ expert) parallelism



UNIVERSITY OF
MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu