Introduction to Parallel Computing (CMSC416 / CMSC818X)

Deep Learning in Parallel

Abhinav Bhatele, Department of Computer Science

UNIVERSITY OF MARYLAND

Announcements

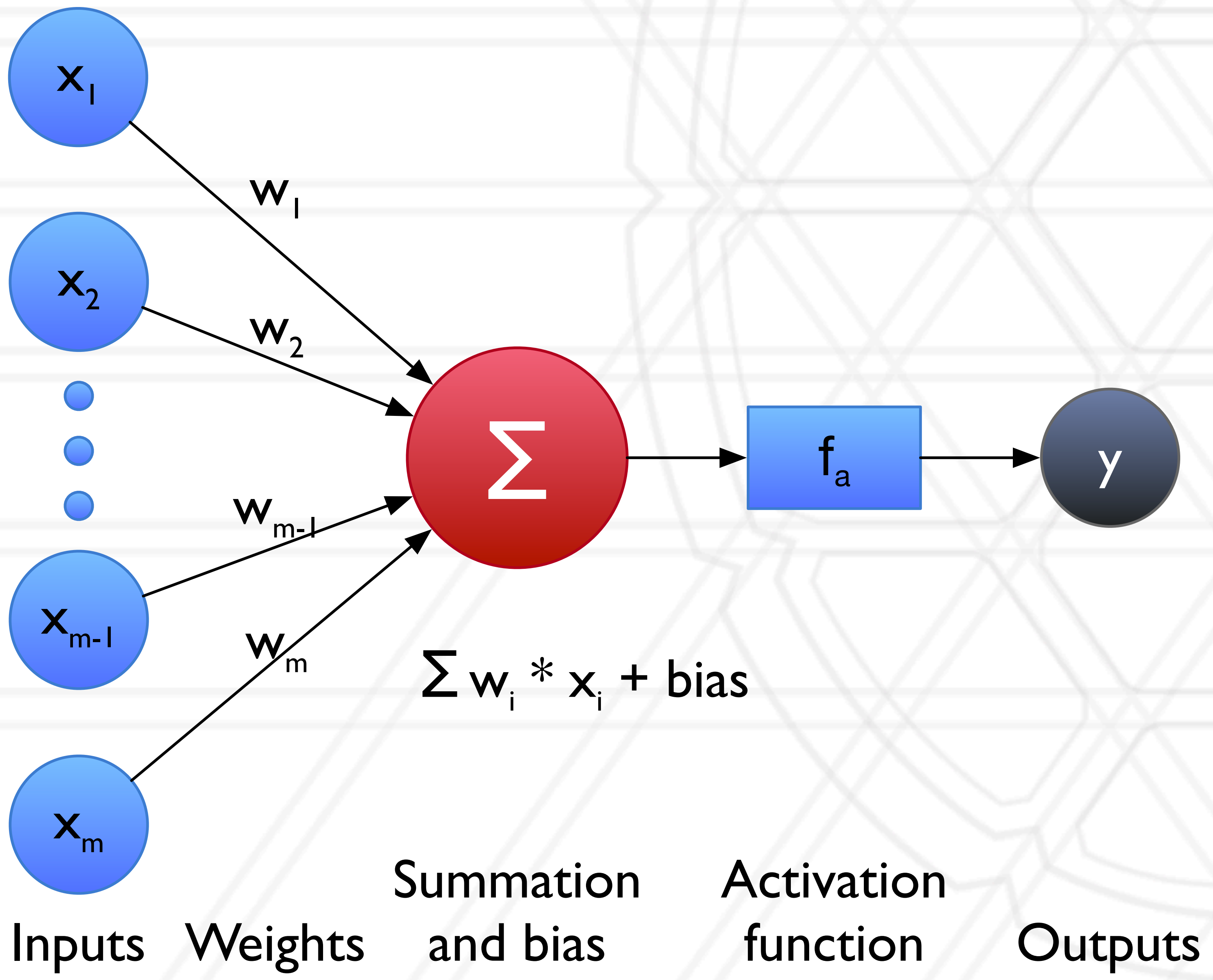
- Extra credit assignment 6 is due on December 7
- Final exam will be posted on gradescope at December 14 12:01 am and will be due on December 14 11:59 pm local time
  - No late submissions allowed
- Course evaluation: https://www.courseevalum.umd.edu

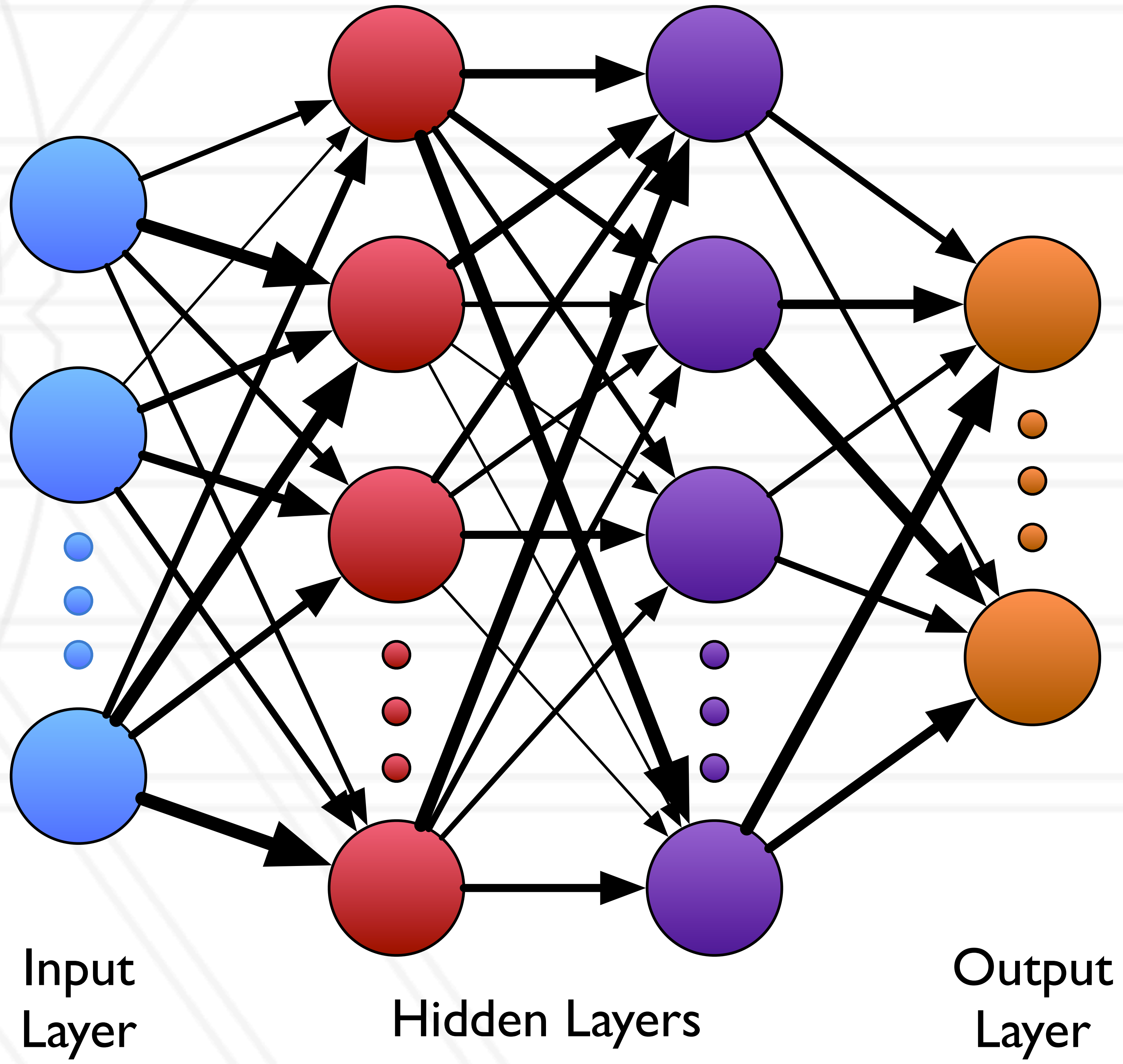
Contact me:

- CMSC416: If you are interested in HPC research
- CMSC818X: If you are interested in collaborating
- If you are an undergrad interested in participating in International Student Cluster Competitions

Deep neural networks

- Neural networks can be used to model complex functions
- Several layers that process “batches” of the input data





Other definitions

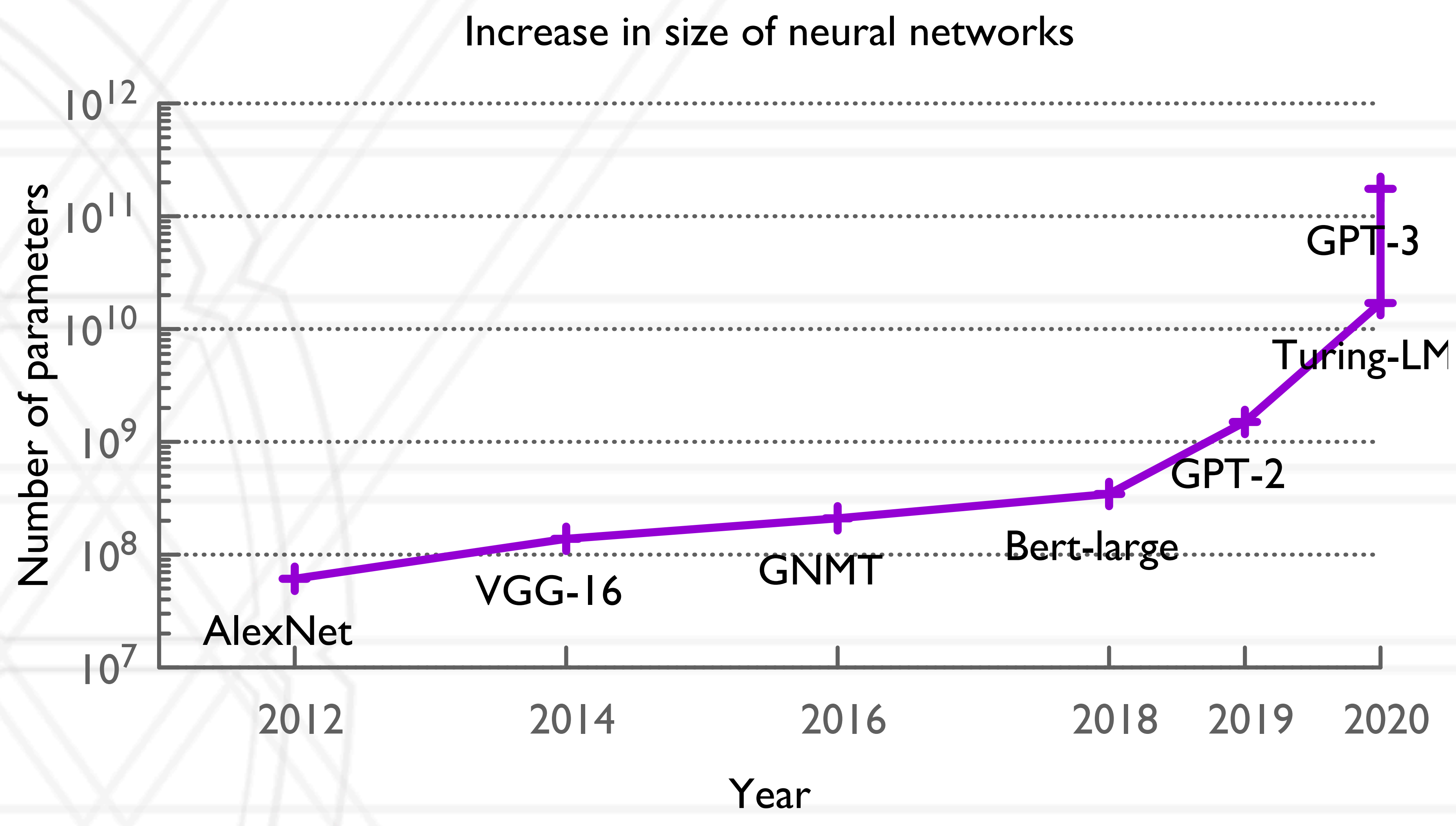
- Learning/training: task of selecting weights that lead to an accurate function
- Loss: a scalar proxy that when minimized leads to higher accuracy
- Gradient descent: process of updating the weights using gradients (derivates) of the loss weighted by a learning rate
- Mini-batch: Small subsets of the dataset processed iteratively
- Epoch: One pass over all the mini-batches

Parallel/distributed training

- Many opportunities for exploiting parallelism
- Iterative process of training (epochs)
- Many iterations per epoch (mini-batches)
- Many layers in DNNs

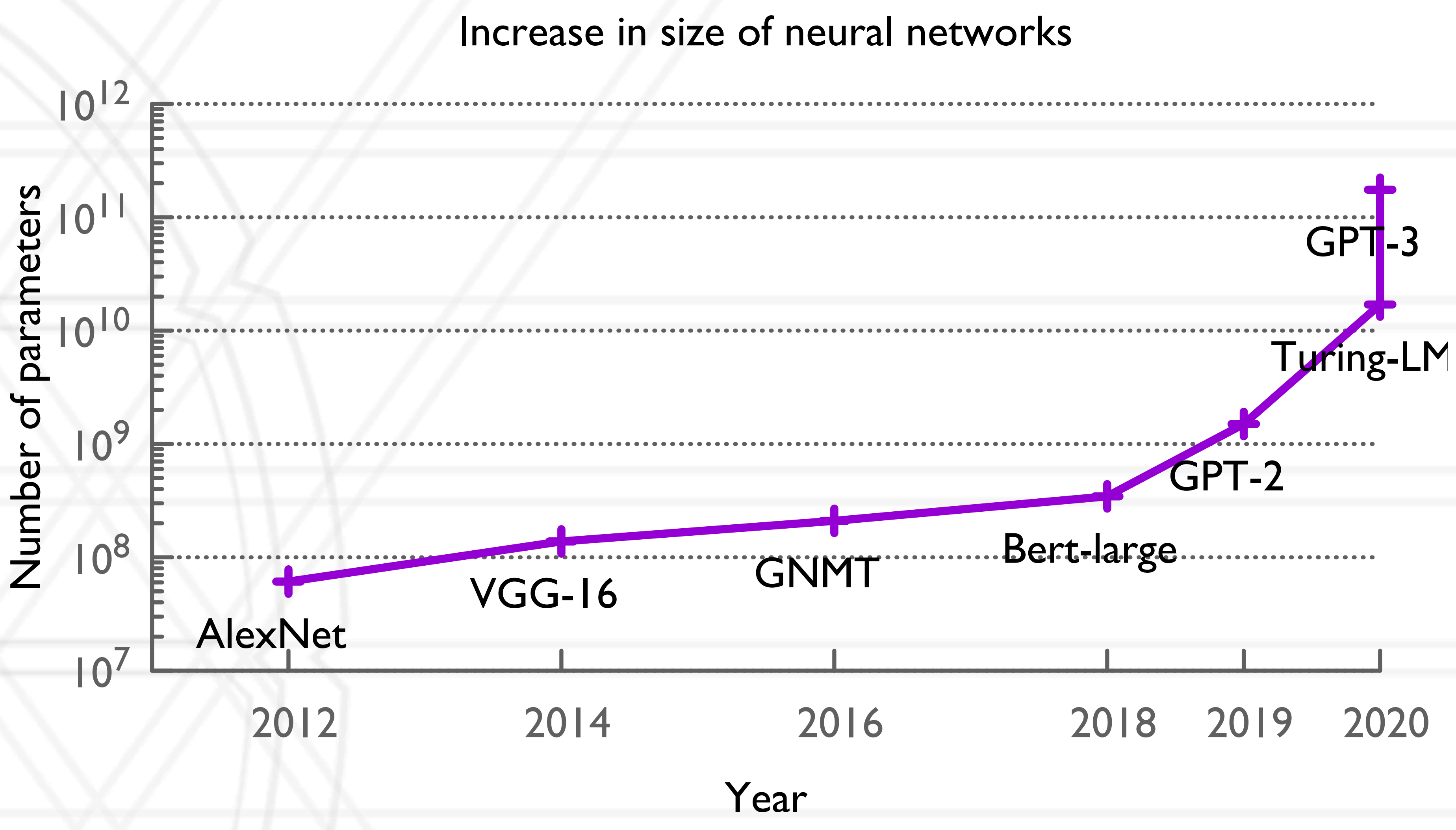
Parallel/distributed training

- Many opportunities for exploiting parallelism
- Iterative process of training (epochs)
- Many iterations per epoch (mini-batches)
- Many layers in DNNs



Parallel/distributed training

- Many opportunities for exploiting parallelism
- Iterative process of training (epochs)
- Many iterations per epoch (mini-batches)
- Many layers in DNNs



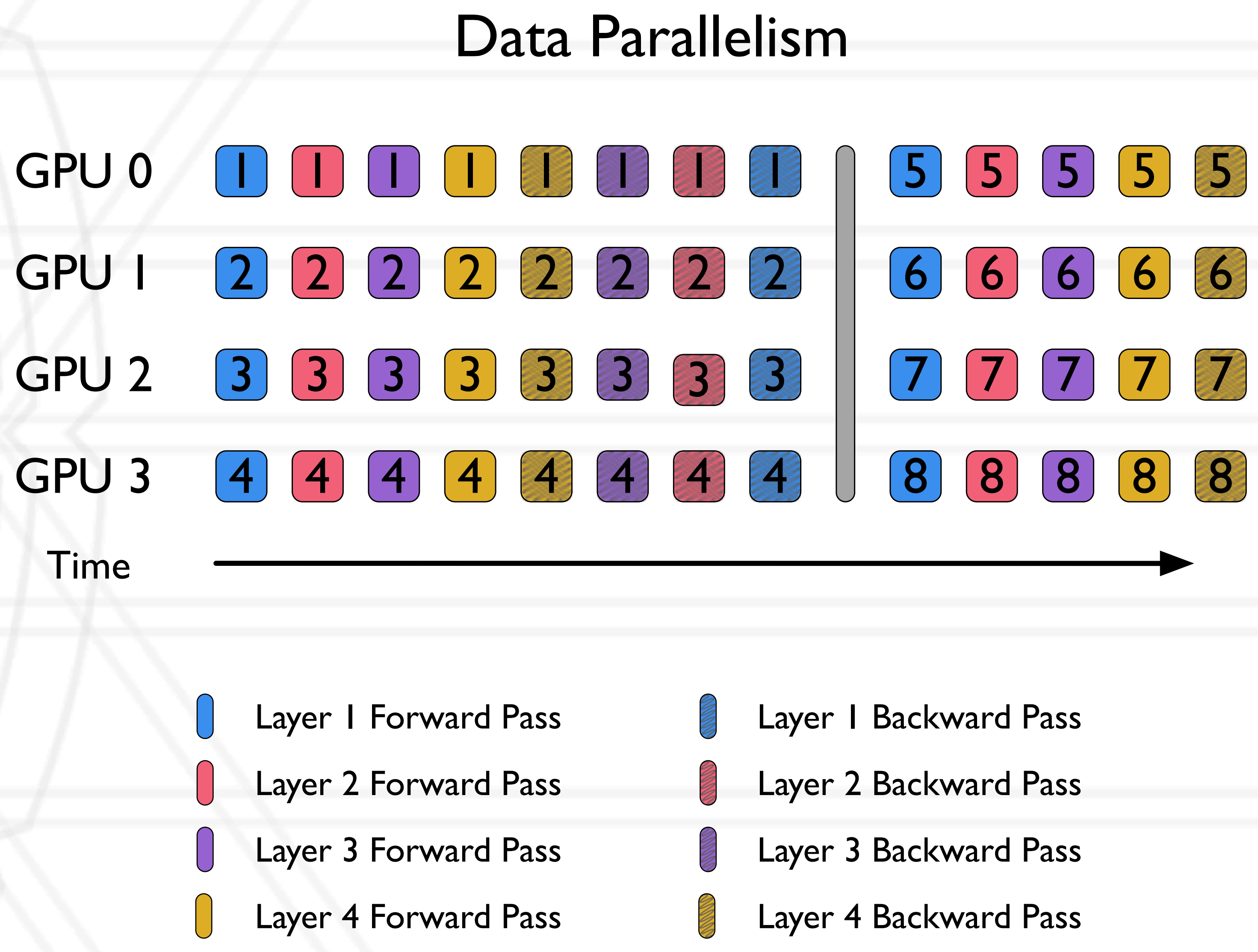
Framework	Type of Parallelism	Largest Accelerator Count	Largest Trained Network (No. of Parameters)
FlexFlow	Hybrid	64 GPUs	24M*
PipeDream	Inter-Layer	16 GPUs	138M
DDP	Data	256 GPUs	345M
GPipe	Inter-Layer	8 GPUs	557M
MeshTensorFlow	Intra-Layer	512-core TPUv2	4.9B
Megatron	Intra-Layer	512 GPUs	8.3B
TorchGPipe	Inter-Layer	8 GPUs	15.8B
KARMA	Data	2048 GPUs	17B
LBANN	Data	3072 CPUs	78.6B
ZeRO	Data	400 GPUs	100B

Data parallelism

- Divide training data among workers (GPUs)
- Each worker has a full copy of the entire NN and processes different mini-batches
- All reduce operation to synchronize gradients

Data parallelism

- Divide training data among workers (GPUs)
- Each worker has a full copy of the entire NN and processes different mini-batches
- All reduce operation to synchronize gradients



Intra-layer parallelism

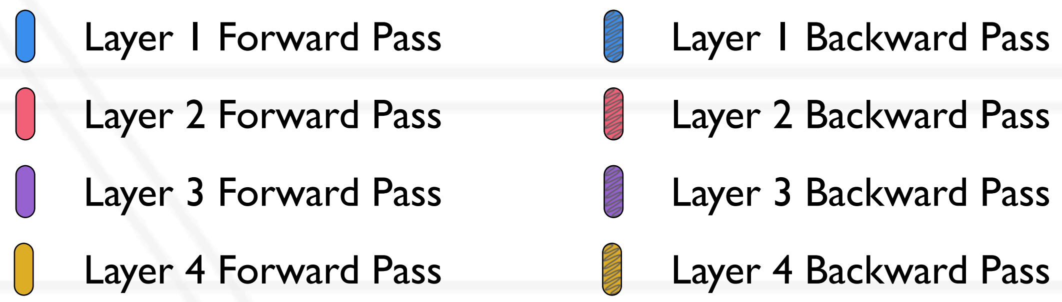
- Enables training neural networks that would not fit on a single GPU
- Distribute the work within a layer between multiple processes/GPUs

Inter-layer parallelism

- Distribute entire layers to different processes/ GPUs
- 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

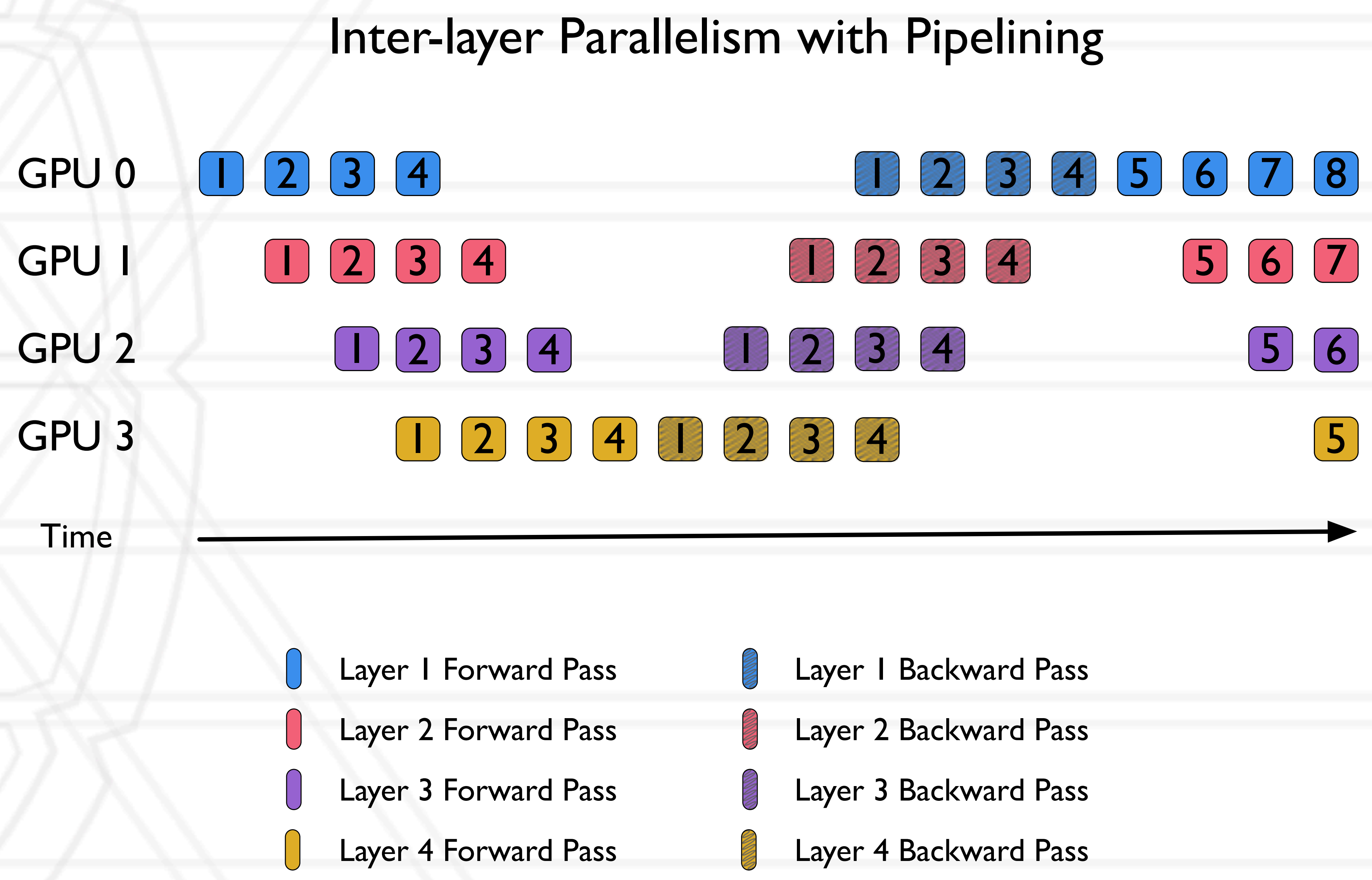
Inter-layer parallelism

- Distribute entire layers to different processes/GPUs
- 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

- Distribute entire layers to different processes/GPUs
- 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



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

Questions?

UNIVERSITY OF MARYLAND

Abhinav Bhatele

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

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