# FUNDAMENTALS OF DEEP LEARNING FOR MULTI-GPUS

## LAB 1 CONCLUSION: DATA AND MODEL PARALLELISM

# DATA PARALLELISM

## Focus of this course

How can we take advantage of multiple GPUs to reduce the training time?

# DATA VS MODEL PARALLELISM

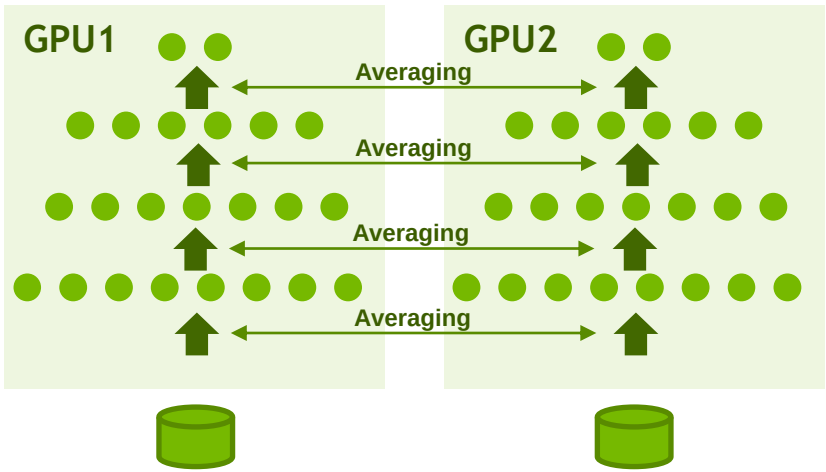
## Comparison

- Data Parallelism
  - Allows you to speed up training
  - All workers train on different data
  - All workers have the same copy of the model
  - Neural network gradients (weight changes) are exchanged
- Model Parallelism
  - Allows you to use a bigger model
  - All workers train on the same data
  - Parts of the model are distributed across GPUs
  - Neural network activations are exchanged

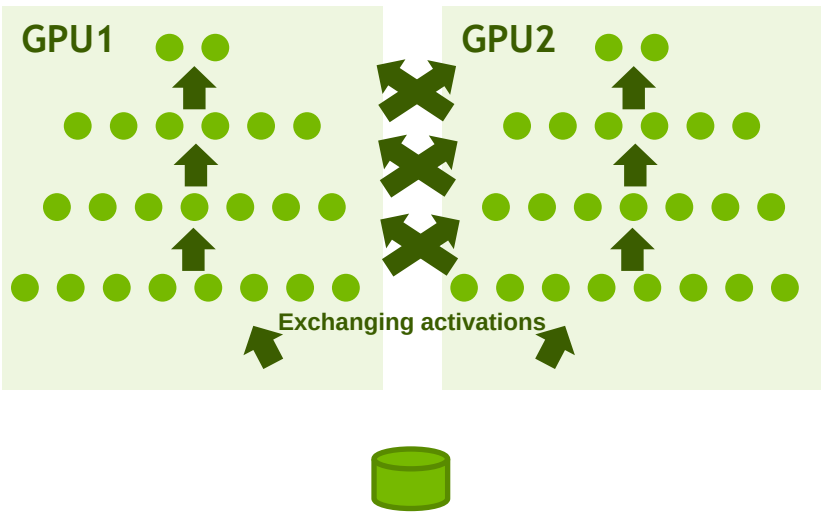
# DATA VS MODEL PARALLELISM

## Comparison

- Data Parallelism



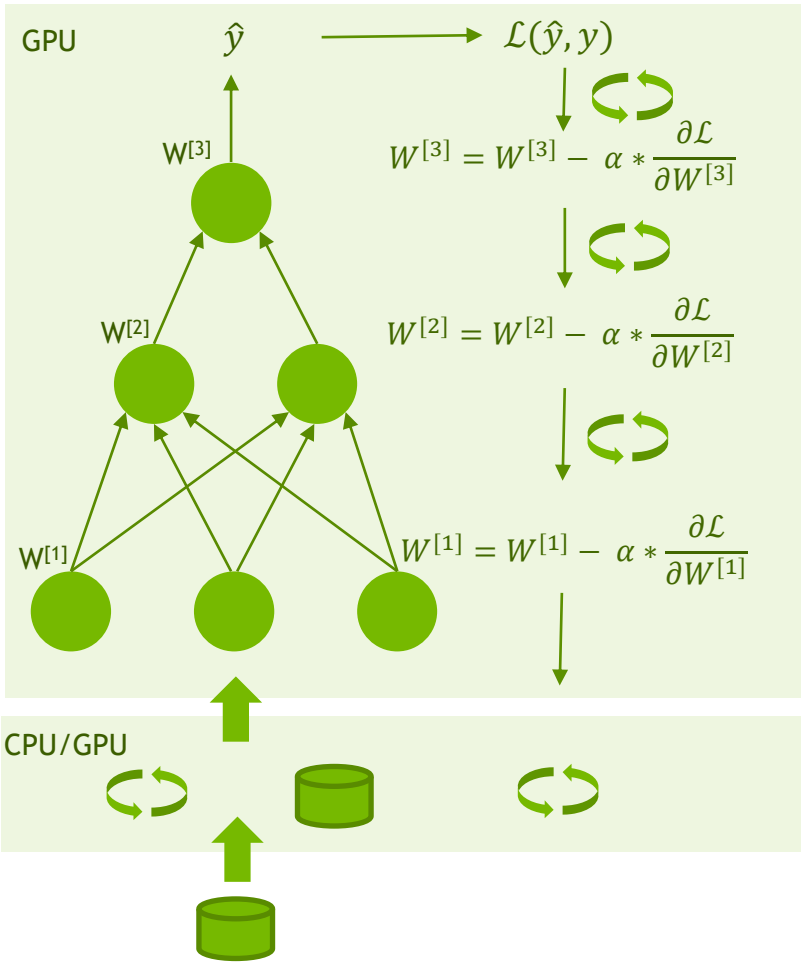

- Model Parallelism

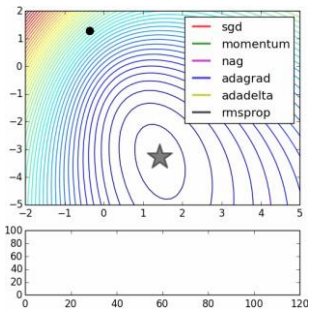



NVIDIA DEEP LEARNING INSTITUTE

# TRAINING A NEURAL NETWORK

## Single GPU



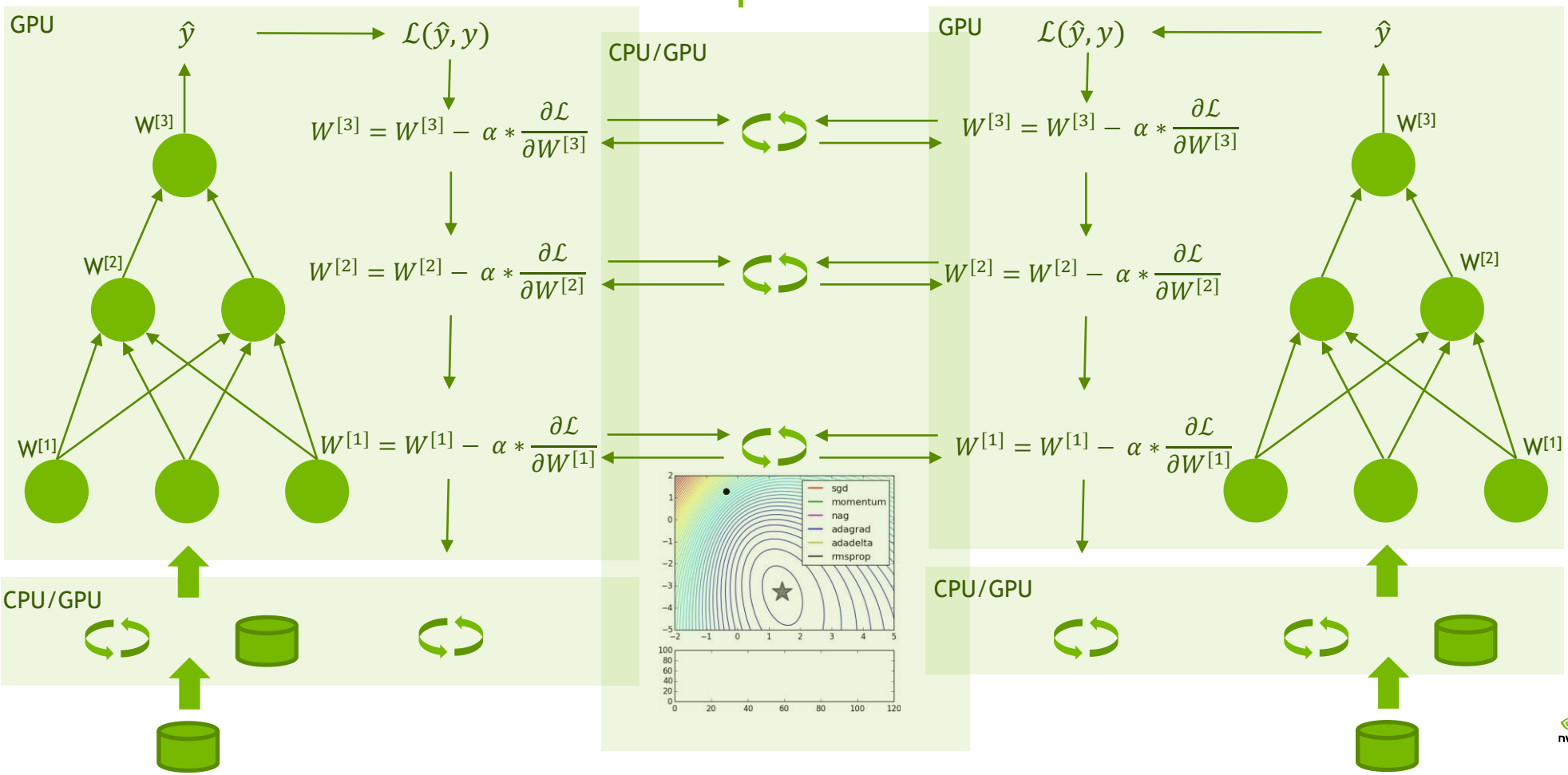




1. Read the data
2. Transport the data
3. Pre-process the data
4. Queue the data
5. Transport the data
6. Calculate activations for layer one
7. Calculate activations for layer two
8. Calculate the output
9. Calculate the loss
10. Backpropagate through layer three
11. Backpropagate through layer two
12. Backpropagate through layer one
13. Execute optimization step
14. Update the weights
15. Return control

NVIDIA DEEP LEARNING INSTITUTE

# TRAINING A NEURAL NETWORK

## Multiple GPUs




NVIDIA DEEP LEARNING INSTITUTE


www.nvidia.com/dli