Caffe2 is ...

- A lightweight framework for deep learning algorithms
- Primarily designed for production use cases
  - Speed is top
  - C++ / Python based interfaces
- First-class distributed support
- Cross-platform
- CV / AR / NLP / Speech / Ranking apps
Optimized for Mobile Inference

Year Class 2015+
- Optimized

Year Class 2013, 2014
- Supported
Training ImageNet using Caffe2 (ResNet-50)

Bring down time to train a new image classifier from multiple days to a single hour

Scale up input batch size to 8192

Use 256 NVIDIA P100 GPUs
(90% scaling efficiency)
Scalable Distributed Training

Caffe2 Trains Up to 7X Faster on a Single DGX-1

- Inception V3: 1 GPU 140 Images/sec, 2 GPUs 509 Images/sec, 4 GPUs 1007 Images/sec, 8 GPUs 1135 Images/sec
- VGG16: 1 GPU 146 Images/sec, 2 GPUs 273 Images/sec, 4 GPUs 571 Images/sec, 8 GPUs 906 Images/sec
- ResNet-50: 1 GPU 231 Images/sec, 2 GPUs 455 Images/sec, 4 GPUs 1785 Images/sec

Caffe2 multi-GPU performance (images/sec) on NVIDIA DGX-1 | Networks: Inception v3, VGG16, ResNet-50 | Batch size: 64 | Number of GPUs: 1, 2, 4, 8
Caffe2 with FP16 Training/Inference

- ~2x training model size & speedups
- Delivering high training throughput on NVIDIA Volta Platforms