Optimizing DNN Operators on Mobile GPUs

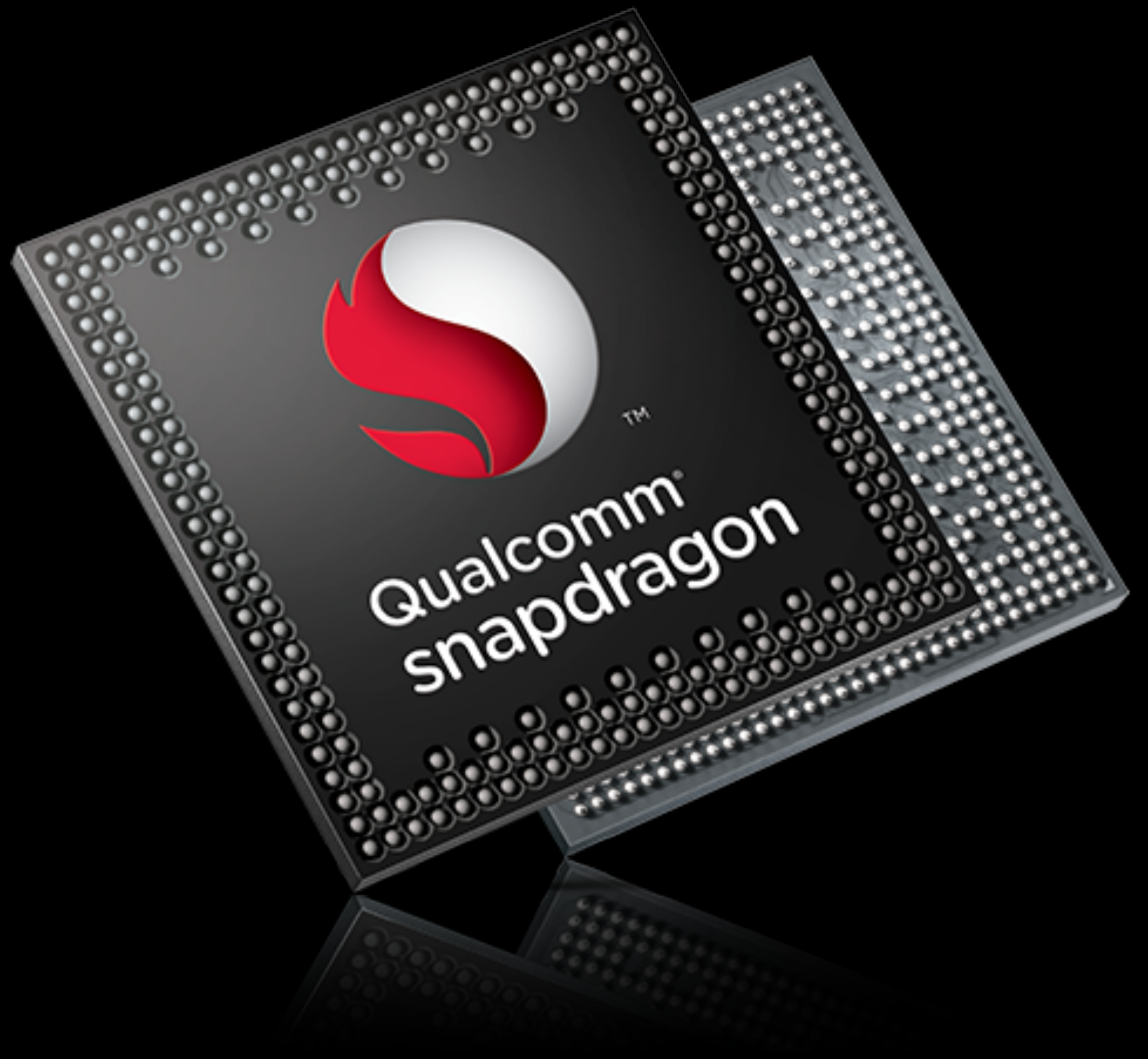
CSC 766 Final Project

Brian Park

DNN Models

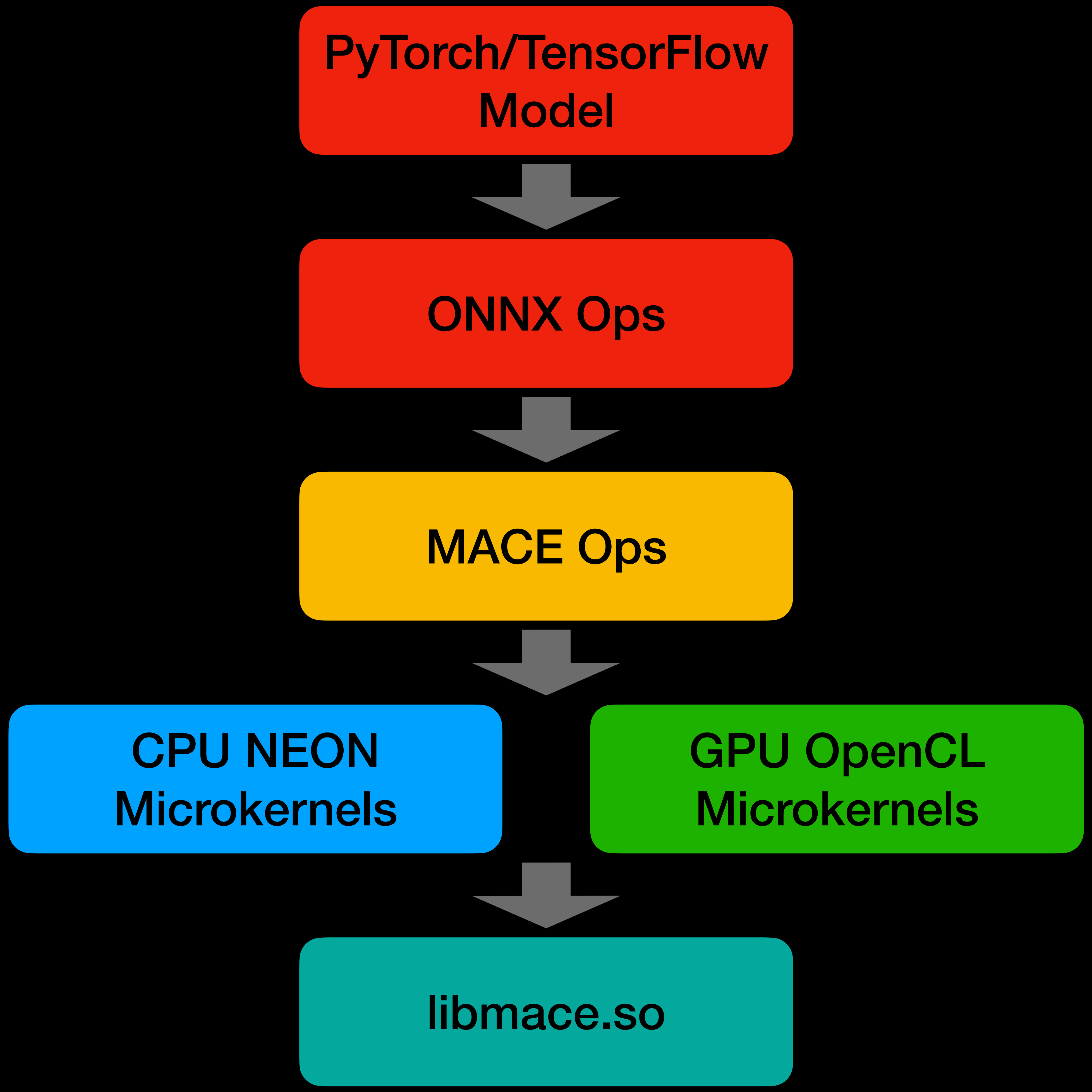
DNN Optimization Task

- ShuffleNet V2+ Small
  - Channel Shuffle
- RegNet (200M)
  - Group Convolution
- Both are models for image classification
- Write GPU kernels in OpenCL for Android GPU to complete support of these DNNs



MACE Framework Overview

- Mobile AI Compute Engine
- Open source library from XiaoMi
- Deep learning inference framework optimized for CPU and GPU on Android platform



PyTorch/TensorFlow Model

ONNX Ops

MACE Ops

CPU NEON Microkernels

GPU OpenCL Microkernels

libmace.so

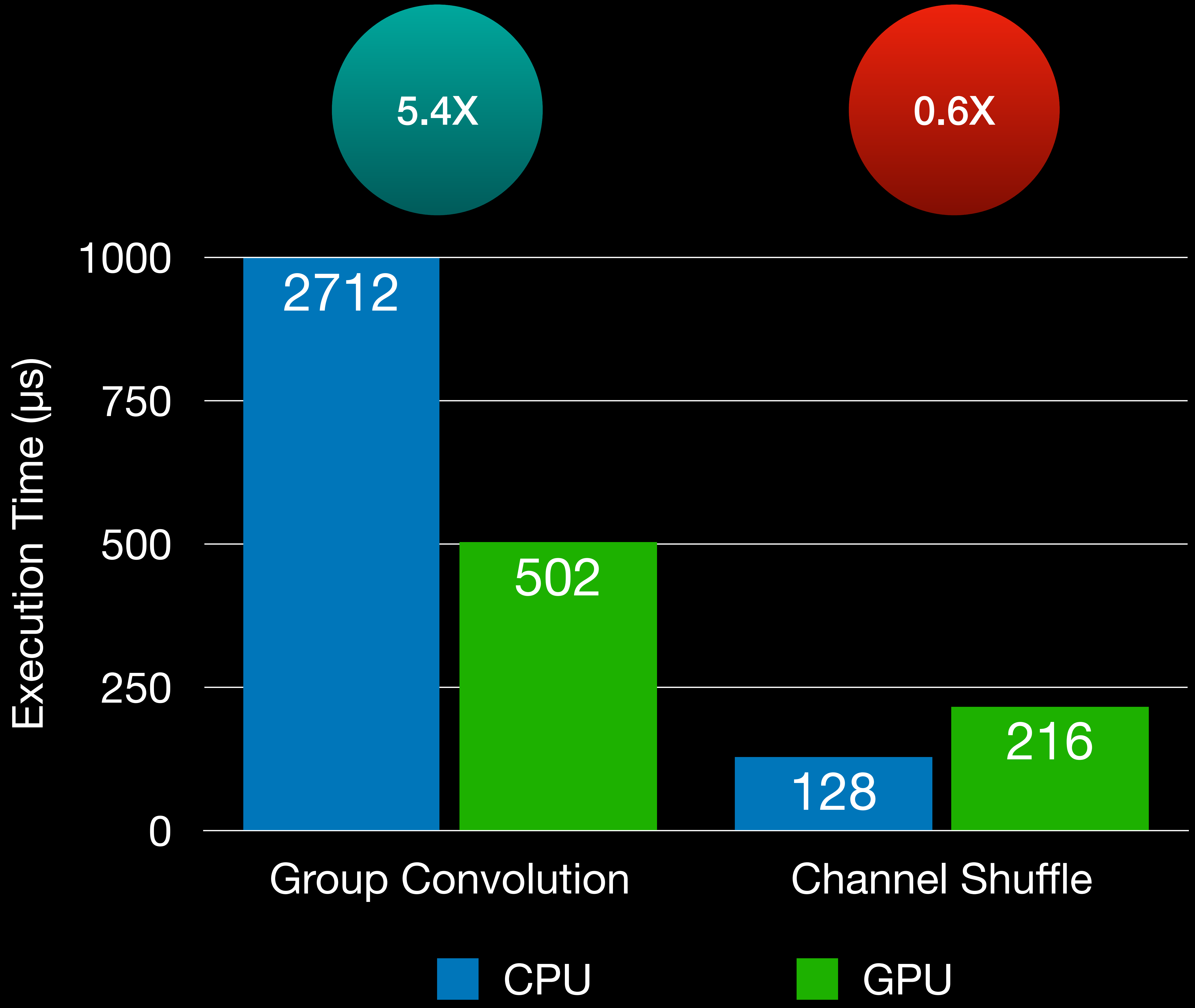
Experimental Results

Device Configuration

- Evaluated on XiaoMi 11 Lite
- Qualcomm SM7150 Snapdragon 732G
- Octa-core CPU (2x2.3 GHz & 6x1.8 GHz)
- Adreno 618 GPU
- Released Spring 2021

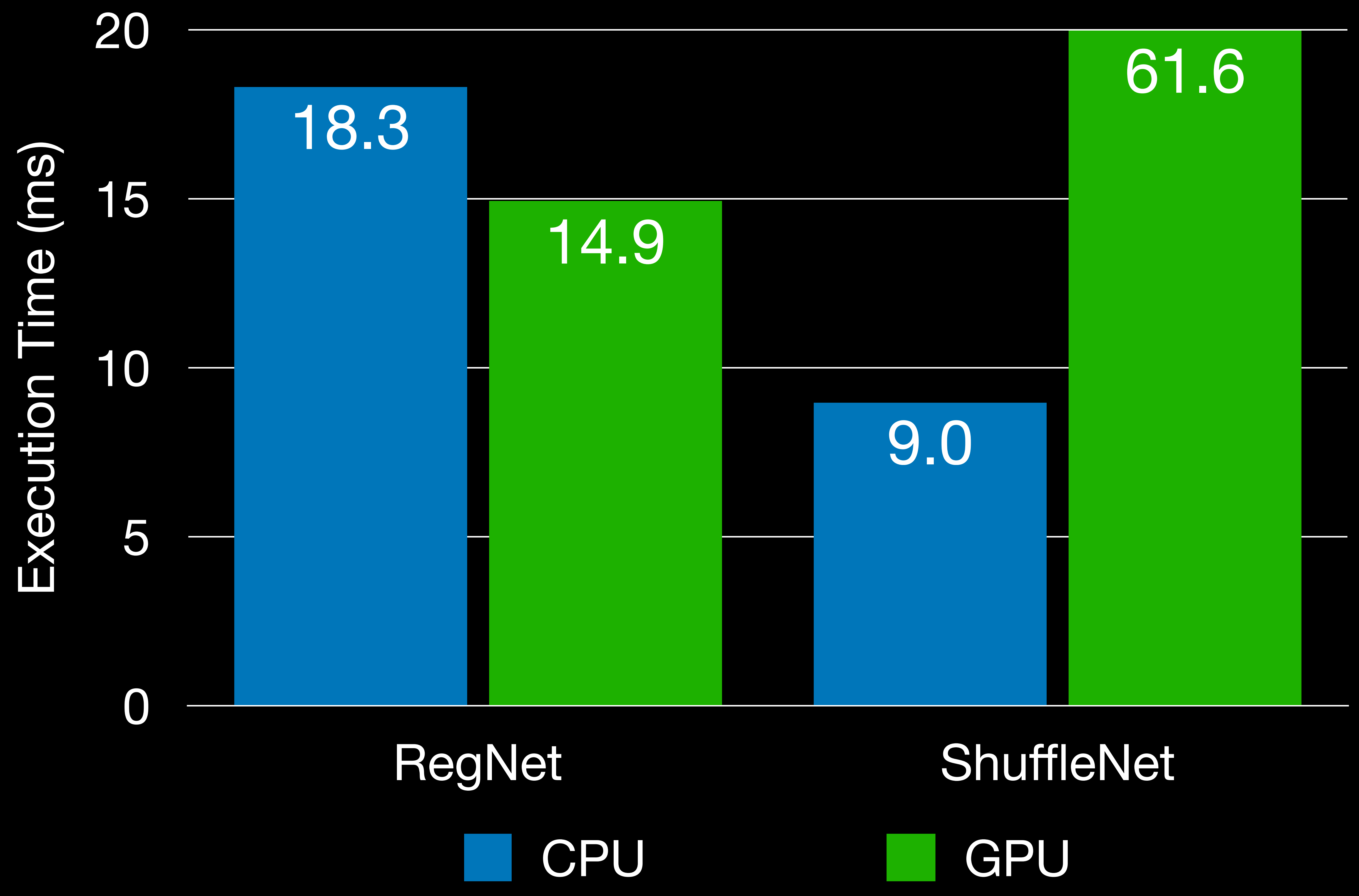
Op Performance

- Group Convolution originally not implemented for both CPU and GPU
- Channel Shuffle supported for CPU and GPU, but needed support for group size of 2 on GPU
  - Channel Shuffle is IO bounded
  - GPU has lower clock frequency compared against CPU



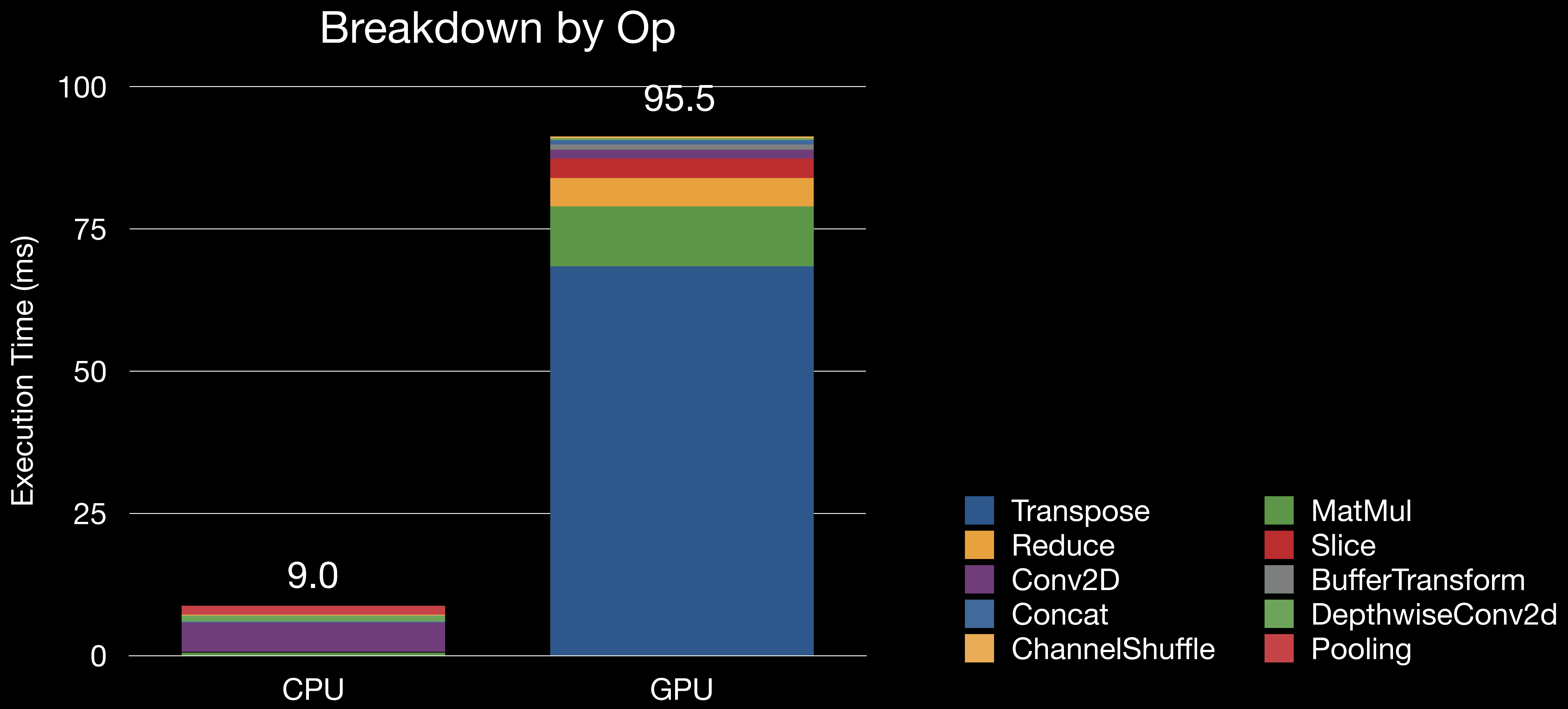
End-to-End Performance of Models

- Batch size of 1
- Parameters for ImageNet
  - Input size: [1, 224, 224, 3]
  - Output size: [1,1000]
- Some ops in GPU configuration fallback to CPU



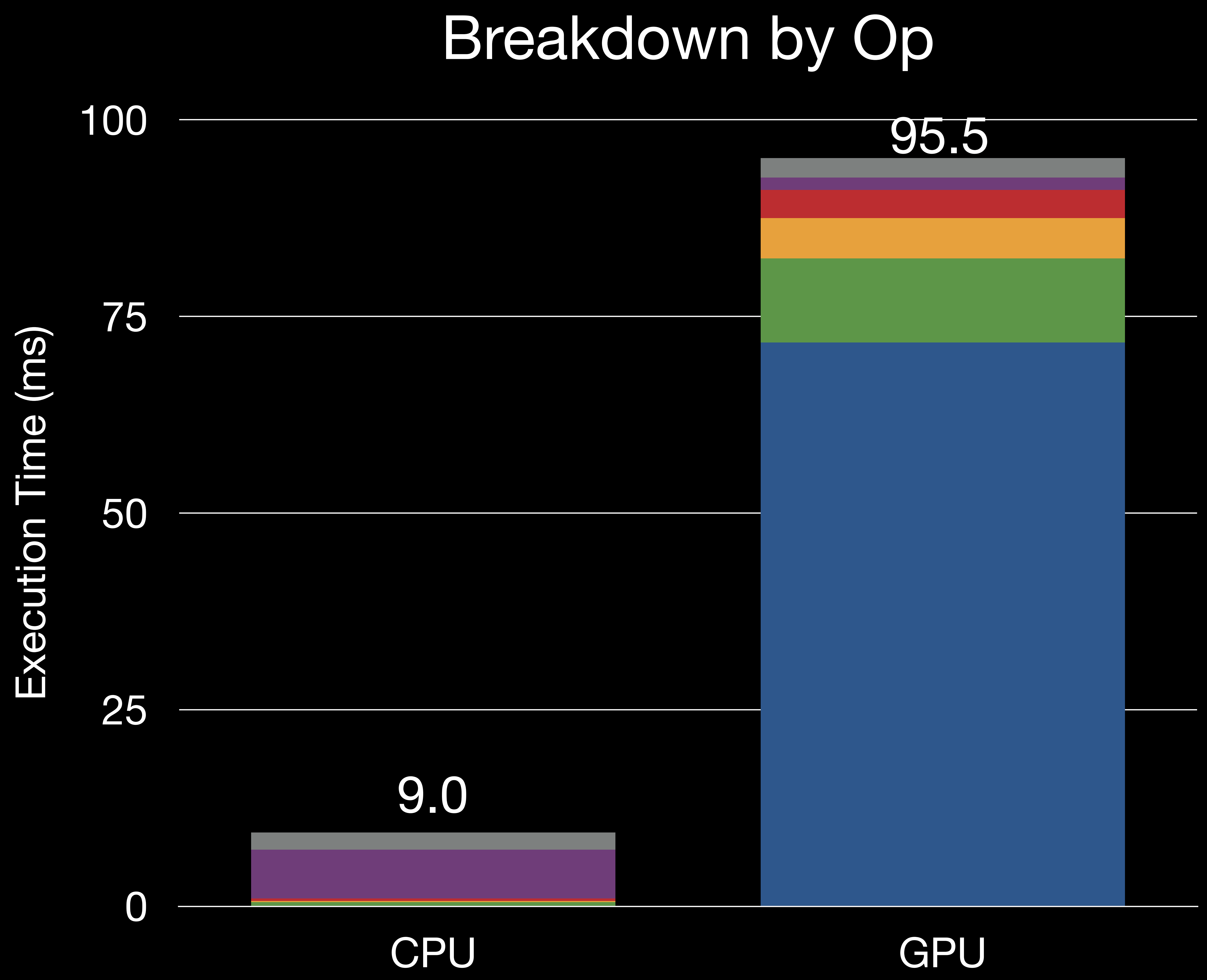
Model Breakdown

ShuffleNet



Model Breakdown

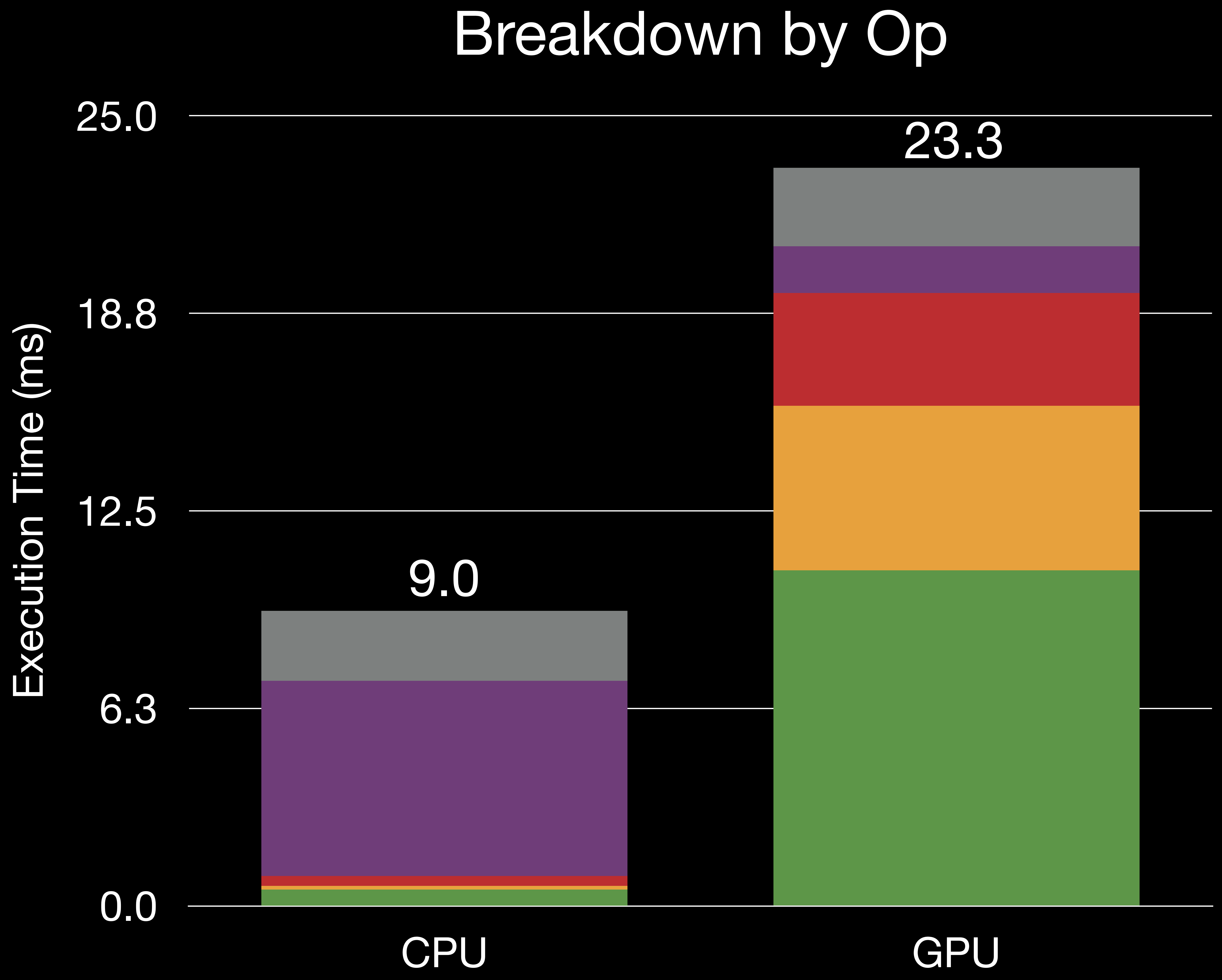
ShuffleNet



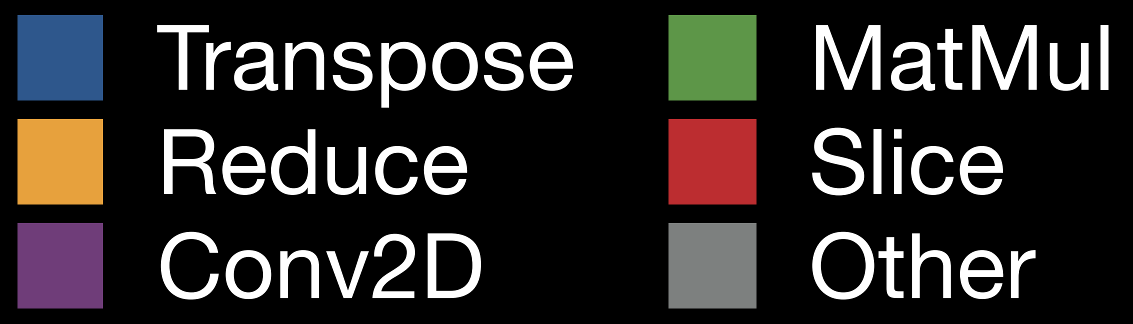
- GPU requires NHWC format
- CPU prefers NCHW format
- GPU Slice and Concat do not support NHWC format

Model Breakdown

Estimated ShuffleNet Performance with Transpose Ops Eliminated



- Need to profile and debug further if MatMul, Slice, and Reduce on GPU configuration can be improved



Challenges and Lessons Learned

- Understanding and using OpenCL for mobile device
- Debugging in Android environment complex
- Lack of active community support as of today
- Combines knowledge from HPC, compiler theory, computer architecture, and ML

Next Steps

- Further improve performance of GPU Kernels
- Eliminate unnecessary transpose ops in ShuffleNet
- Optimize CPU performance to complete support
- Create a Pull Request if all things go well

Thank You!

Acknowledgements

- CSC 766
- Dr. Shen
- Jiexiong Guan