Inference Times for CNN with different optimizations and compressions 3444.69 Base PyTorch (No optimization) 3500 ONNX (No optimization) 3000 Optimization + Compression (oX cY) Inference Time (s) 2500 2000 1516.26 1500 1000 663.07 666.16 680.57 671.57 665.25 667.36 476.48 500 262.79 pase lation pase in mit lation or intinio or 100 to 00 to