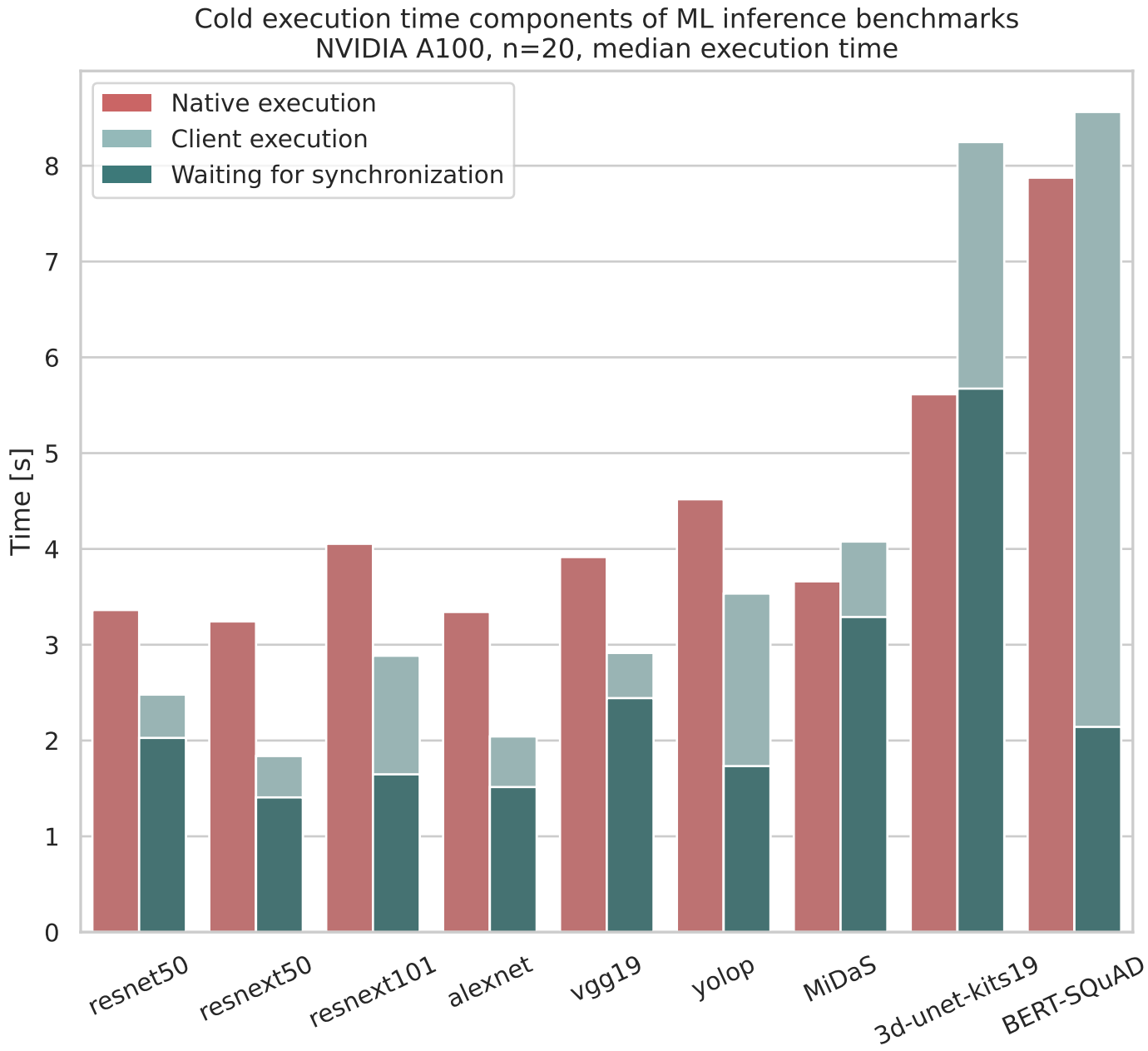
Cold execution time components of ML inference benchmarks
NVIDIA A100, n=20, median execution time
Native execution
Client execution
Waiting for synchronization
Time [s]
0
1
2
3
4
5
6
7
8
resnet50
resnext50
resnext101
alexnet
vgg19
yolop
MiDaS
3d-unet-kits19
BERT-SQuAD