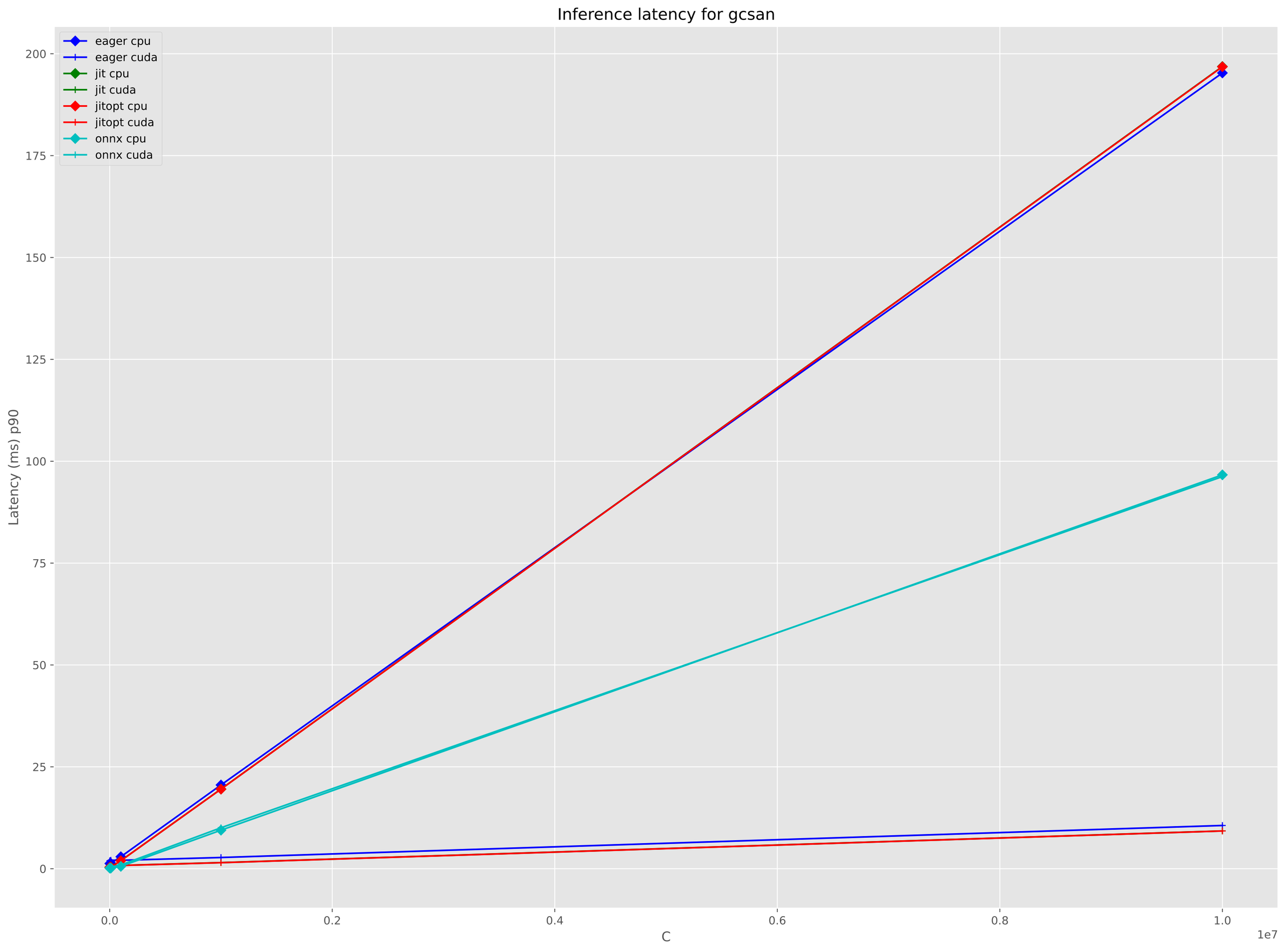
Inference latency for gcsan
eager cpu
eager cuda
jit cpu
jit cuda
jitopt cpu
jitopt cuda
onnx cpu
onnx cuda
Latency (ms) p90
200
175
150
125
100
75
50
25
0
0.0
0.2
0.4
0.6
0.8
1.0
1e7
C