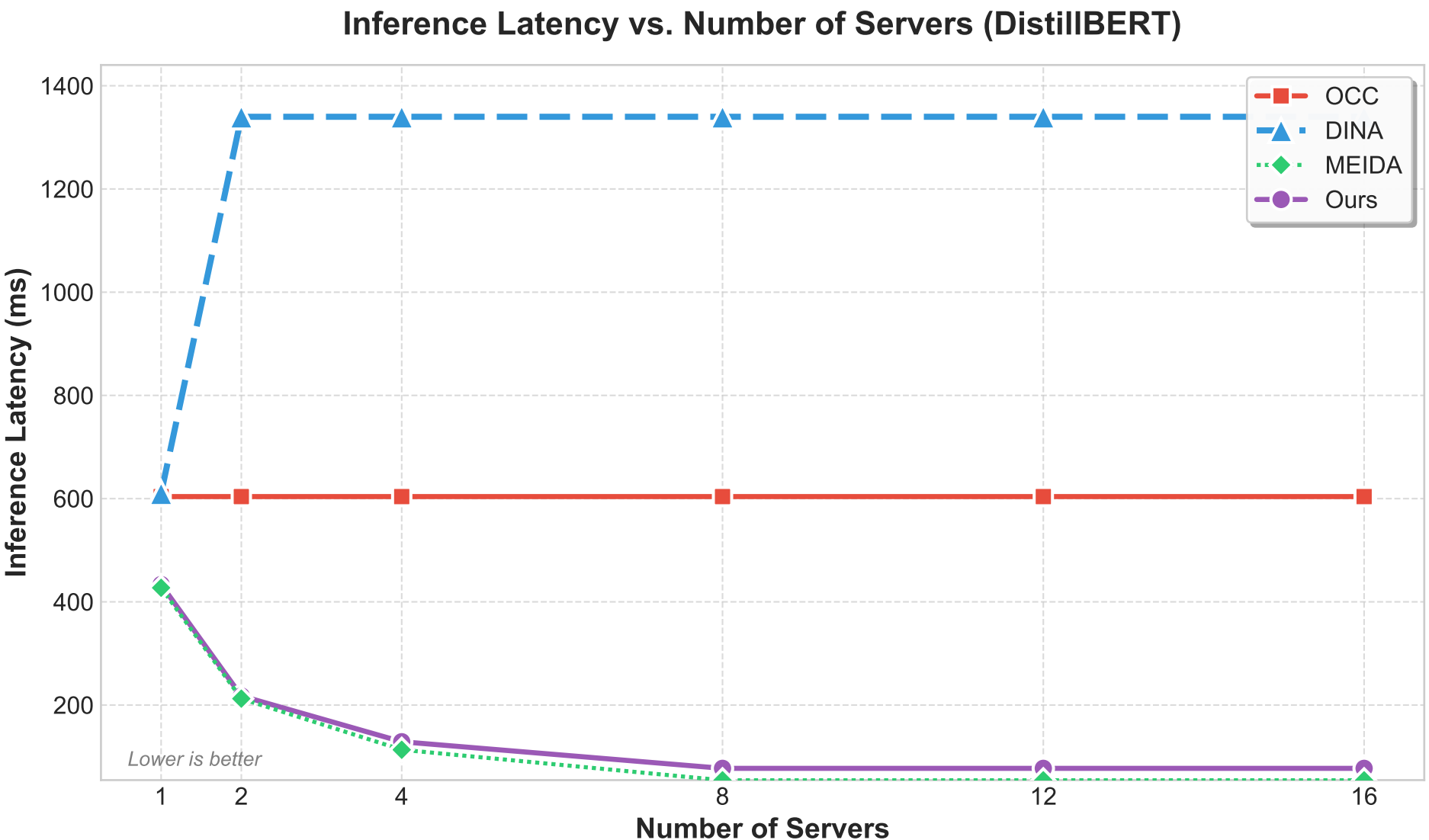
Inference Latency vs. Number of Servers (DistillBERT)
Inference Latency (ms)
1400
1200
1000
800
600
400
200
Lower is better
1
2
4
8
12
16
Number of Servers
OCC
DINA
MEIDA
Ours