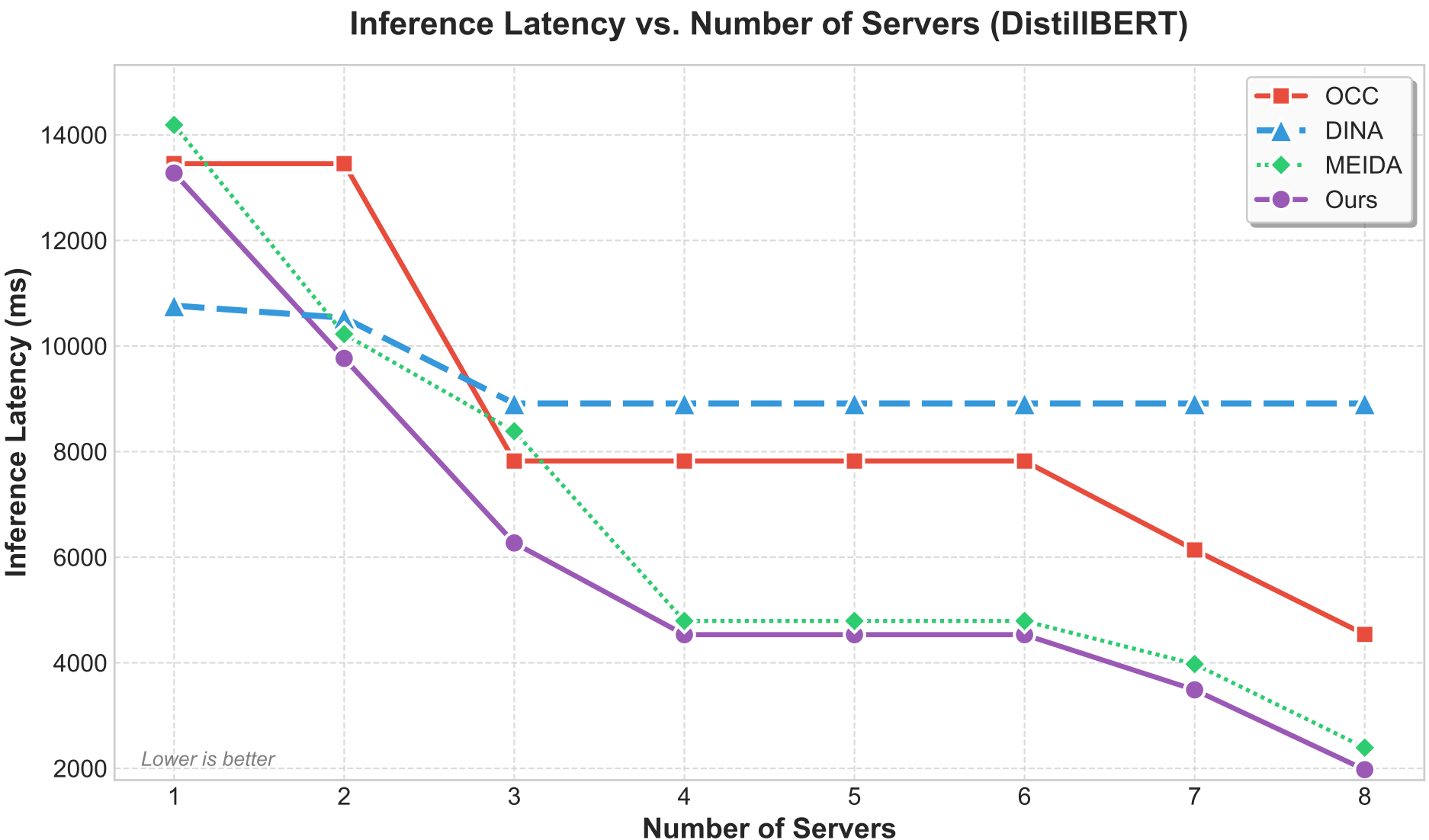
Inference Latency vs. Number of Servers (DistillBERT)
Inference Latency (ms)
14000
12000
10000
8000
6000
4000
2000
1
2
3
4
5
6
7
8
Number of Servers
Lower is better
OCC
DINA
MEIDA
Ours