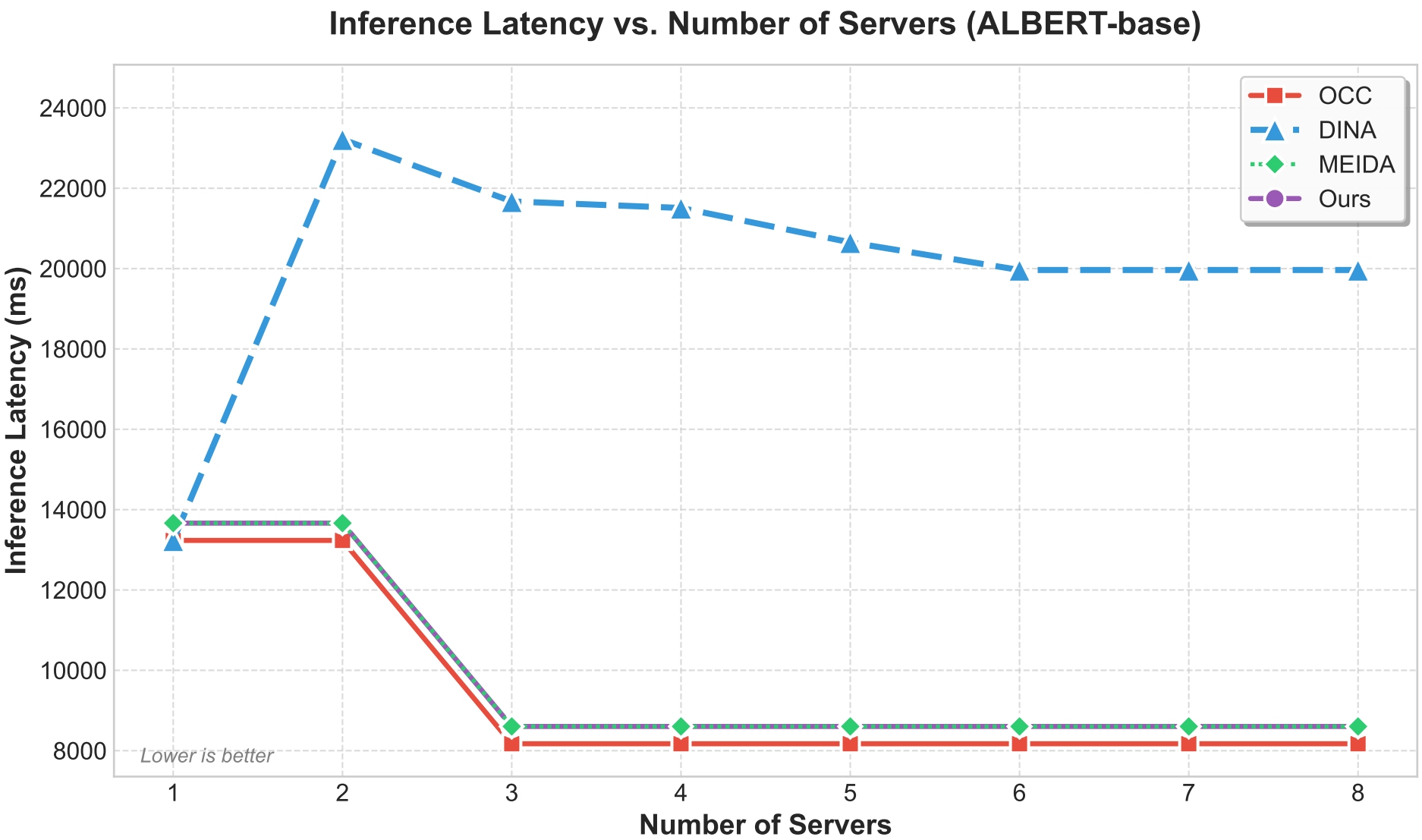
Inference Latency vs. Number of Servers (ALBERT-base)
Inference Latency (ms)
24000
22000
20000
18000
16000
14000
12000
10000
8000
Lower is better
1
2
3
4
5
6
7
8
Number of Servers
OCC
DINA
MEIDA
Ours