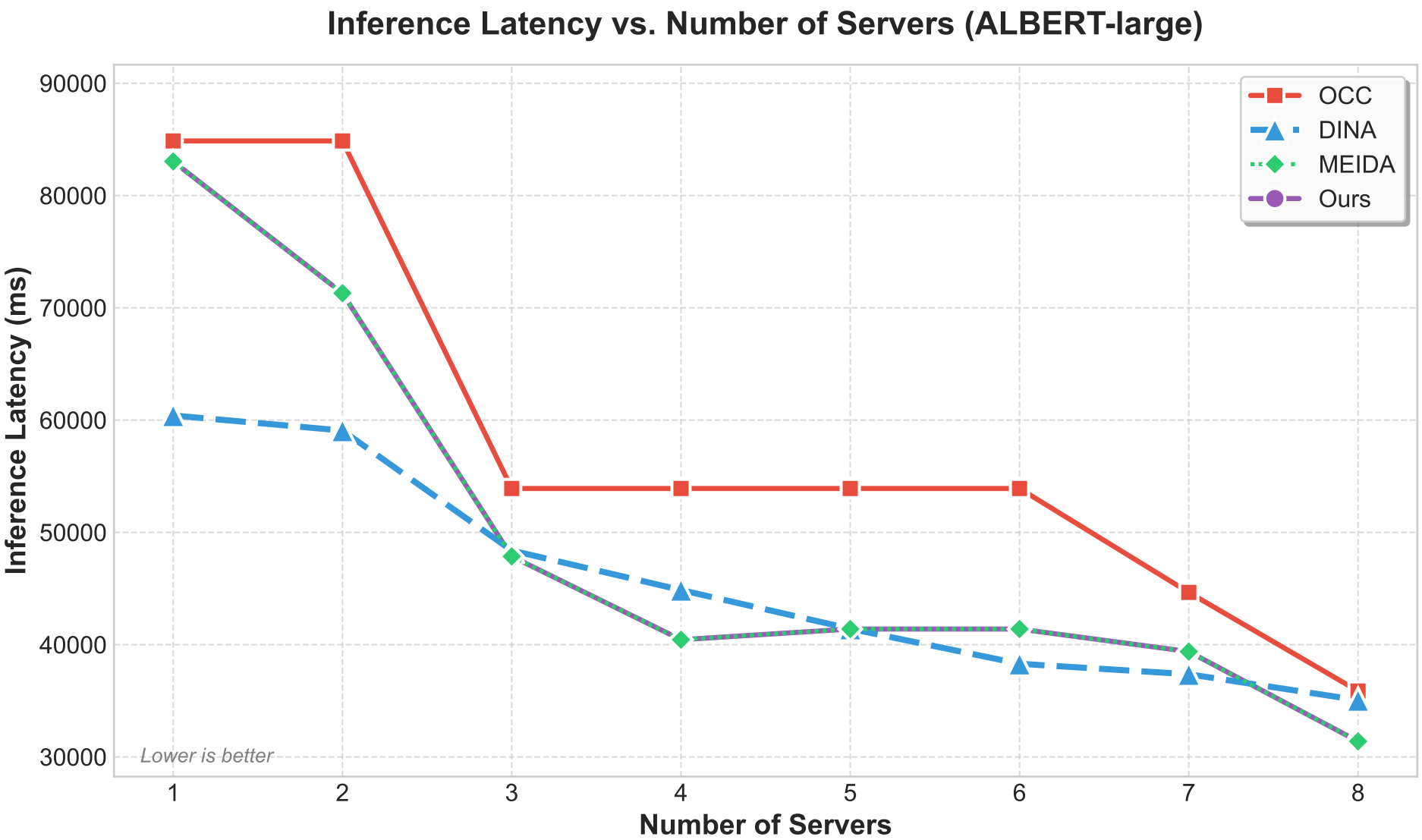
Inference Latency vs. Number of Servers (ALBERT-large)
Inference Latency (ms)
90000
80000
70000
60000
50000
40000
30000
1
2
3
4
5
6
7
8
Number of Servers
OCC
DINA
MEIDA
Ours
Lower is better