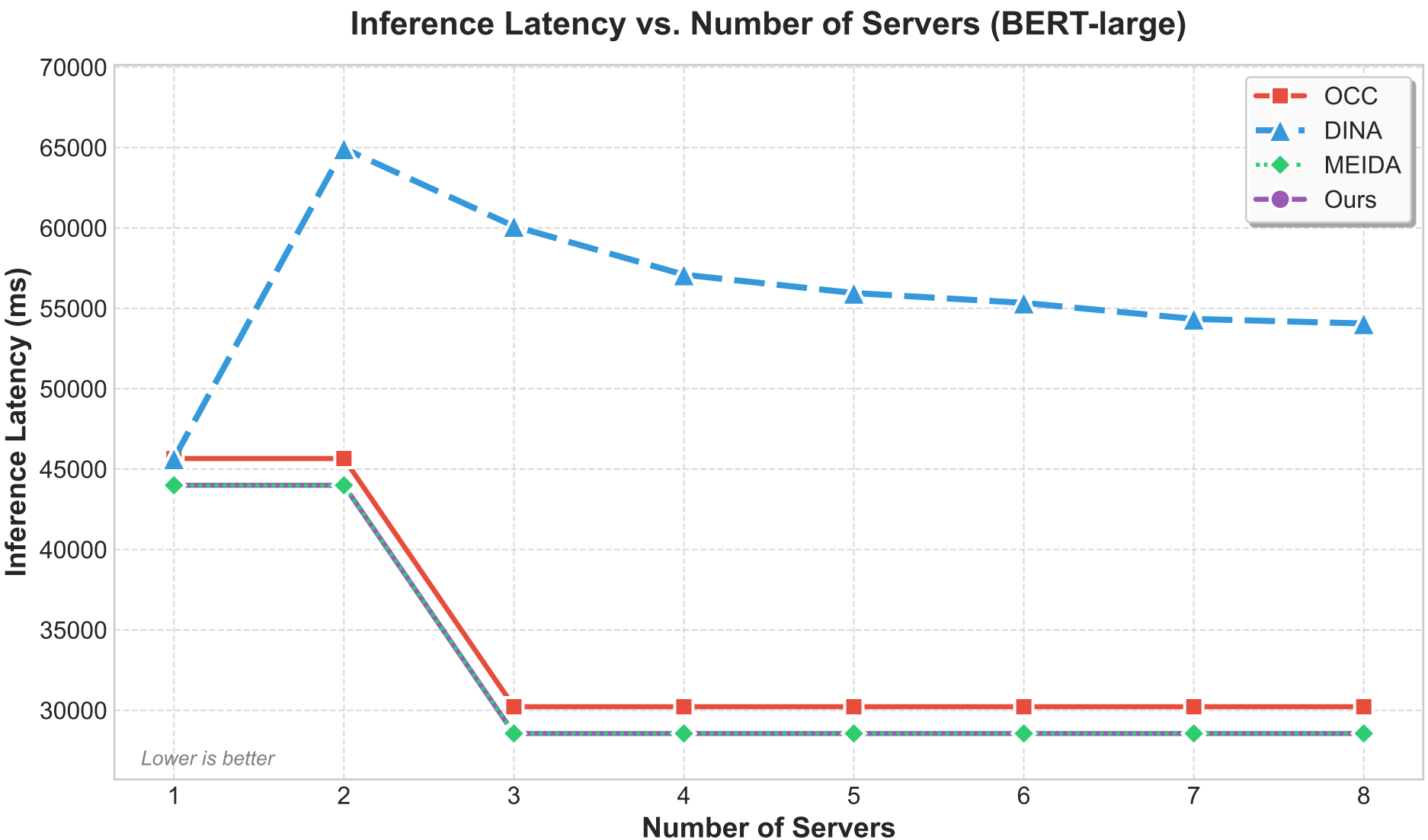
Inference Latency vs. Number of Servers (BERT-large)
Inference Latency (ms)
70000
65000
60000
55000
50000
45000
40000
35000
30000
Lower is better
1
2
3
4
5
6
7
8
Number of Servers
OCC
DINA
MEIDA
Ours