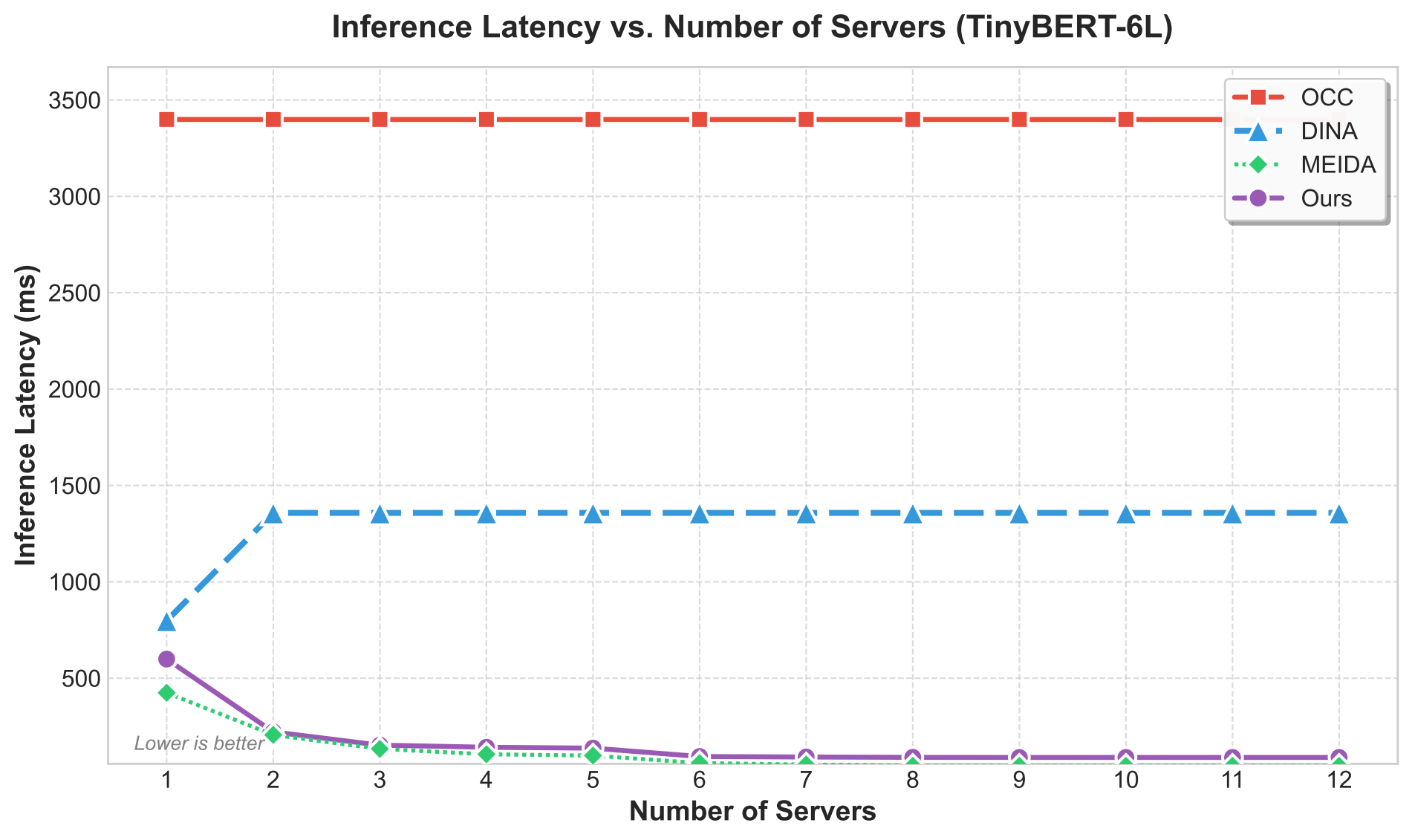
Inference Latency vs. Number of Servers (TinyBERT-6L)
Inference Latency (ms)
3500
3000
2500
2000
1500
1000
500
Lower is better
1
2
3
4
5
6
7
8
9
10
11
12
Number of Servers
OCC
DINA
MEIDA
Ours