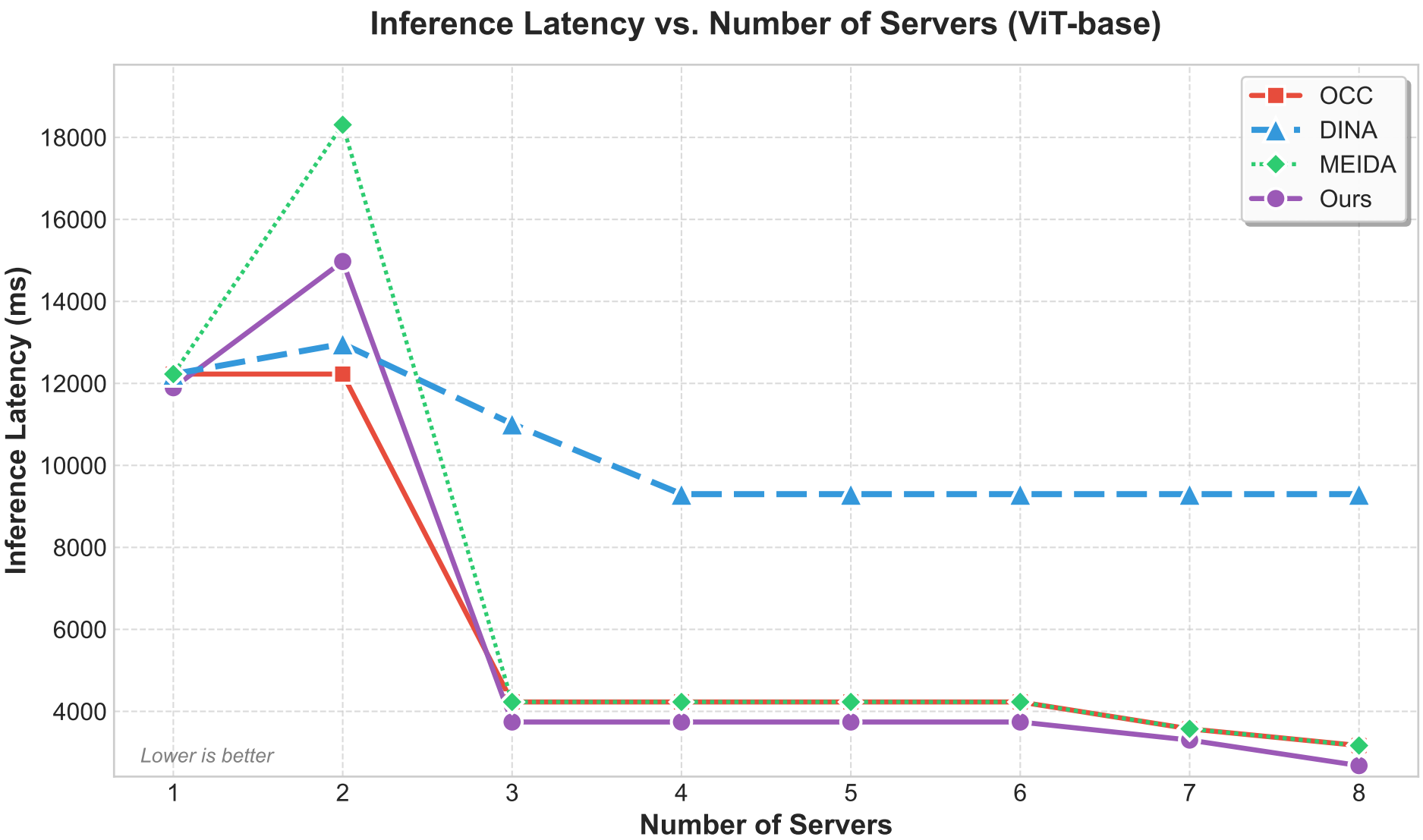
Inference Latency vs. Number of Servers (ViT-base)
Inference Latency (ms)
Number of Servers
OCC
DINA
MEIDA
Ours
Lower is better
18000
16000
14000
12000
10000
8000
6000
4000
1
2
3
4
5
6
7
8