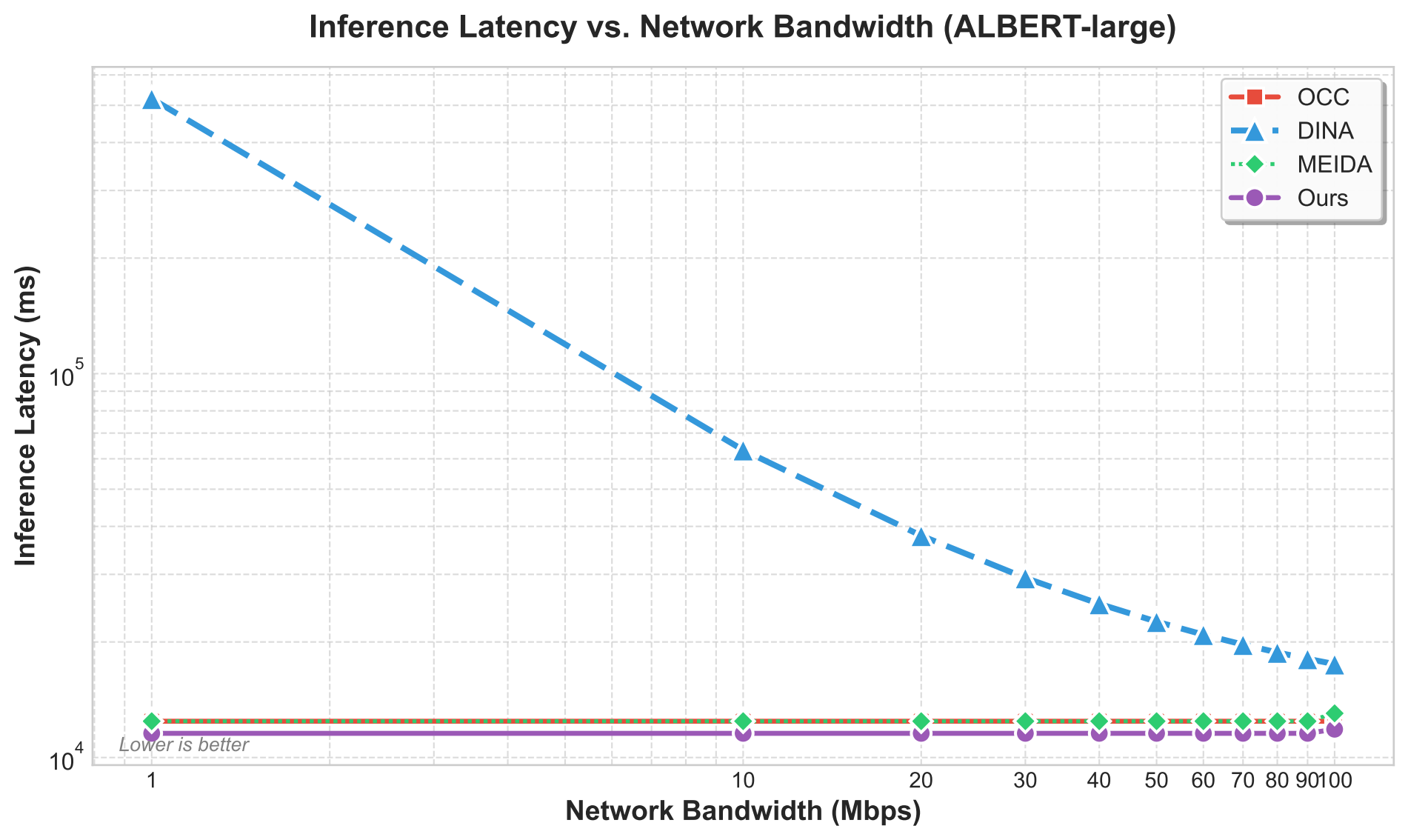
Inference Latency vs. Network Bandwidth (ALBERT-large)
OCC
DINA
MEIDA
Ours
Inference Latency (ms)
10^5
10^4
Lower is better
1
10
20
30
40
50
60
70
80
90
100
Network Bandwidth (Mbps)