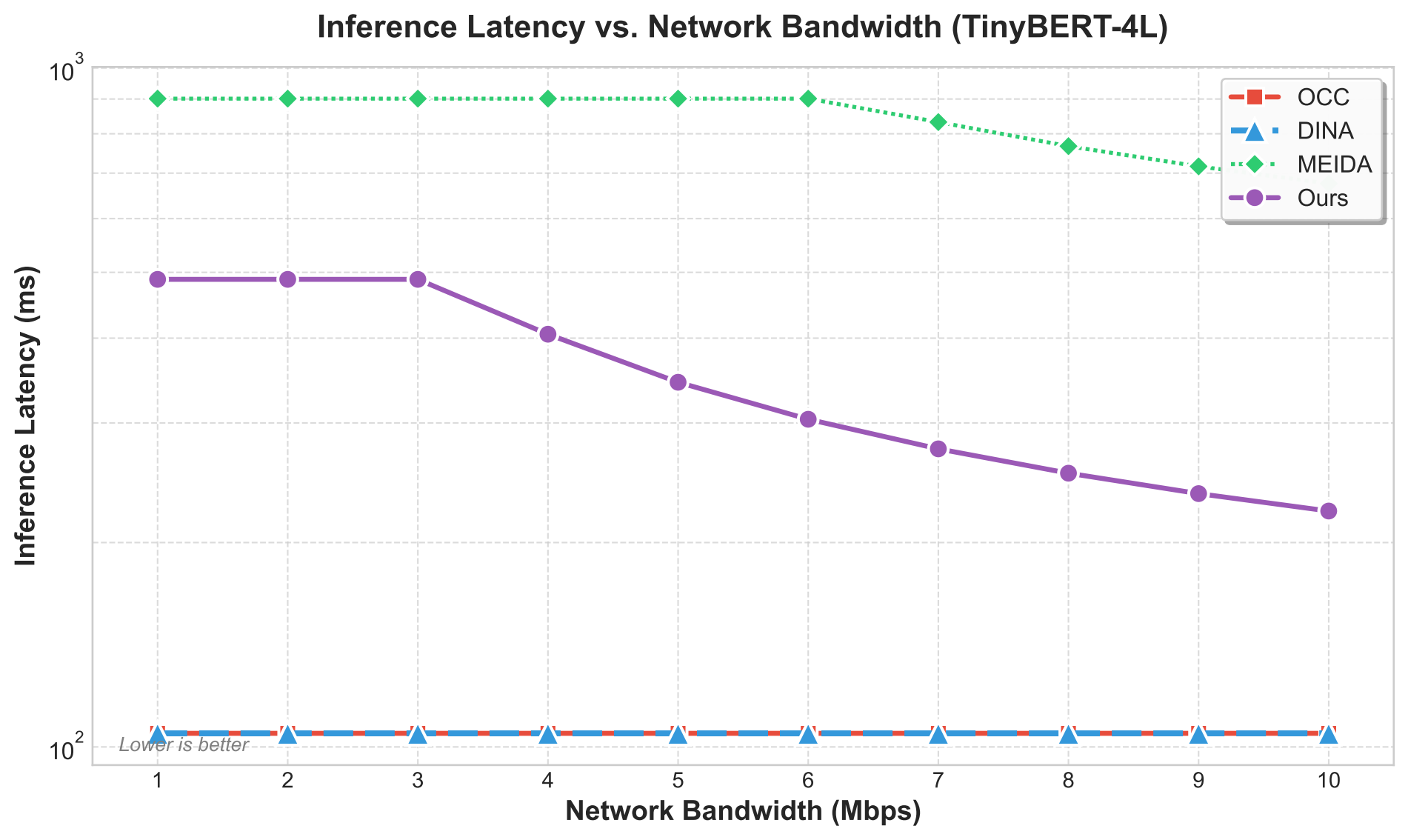
Inference Latency vs. Network Bandwidth (TinyBERT-4L)
Inference Latency (ms)
10^3
10^2
Lower is better
1
2
3
4
5
6
7
8
9
10
Network Bandwidth (Mbps)
OCC
DINA
MEIDA
Ours