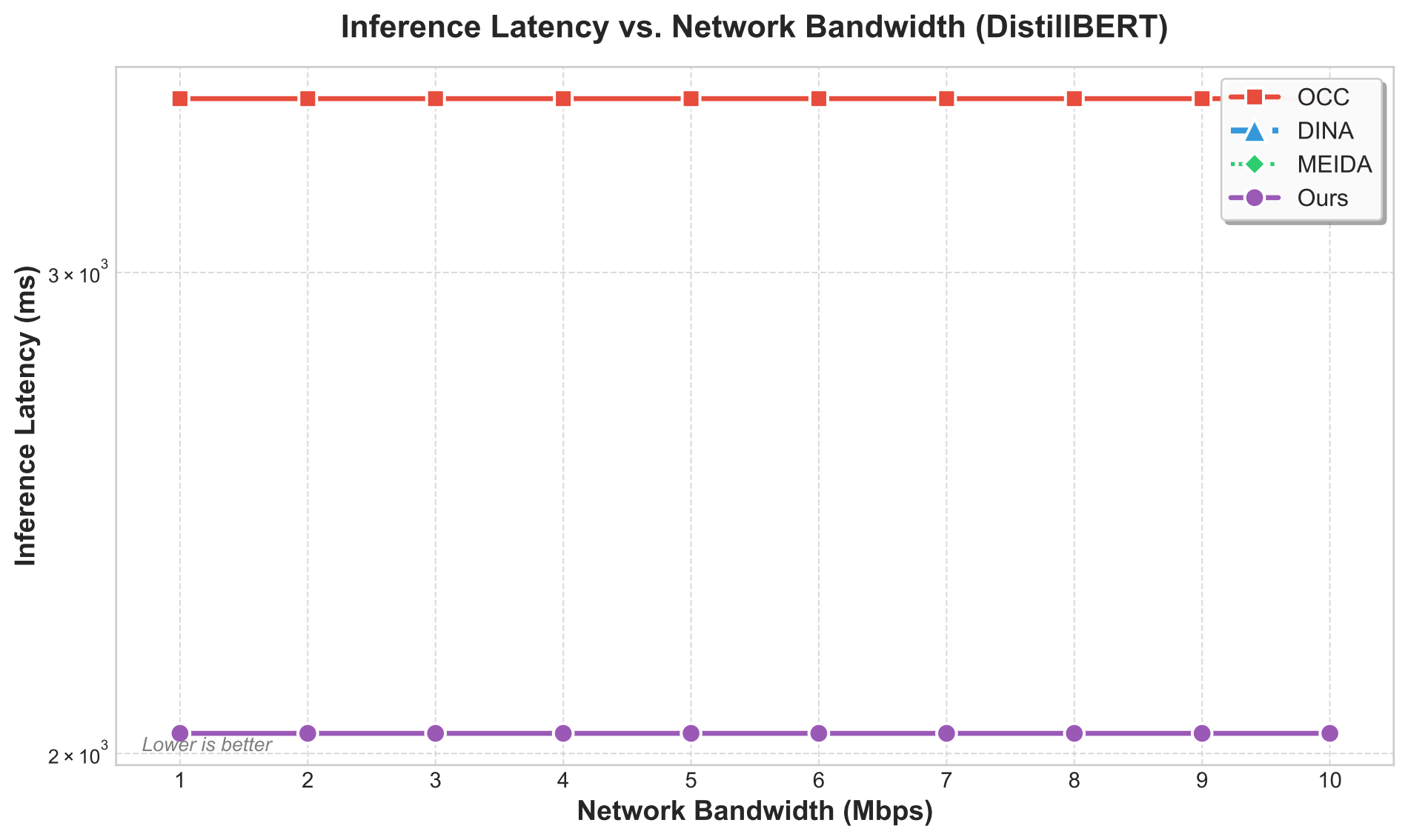
Inference Latency vs. Network Bandwidth (DistillBERT)
Inference Latency (ms)
3×10^3
2×10^3
Lower is better
1
2
3
4
5
6
7
8
9
10
Network Bandwidth (Mbps)
OCC
DINA
MEIDA
Ours