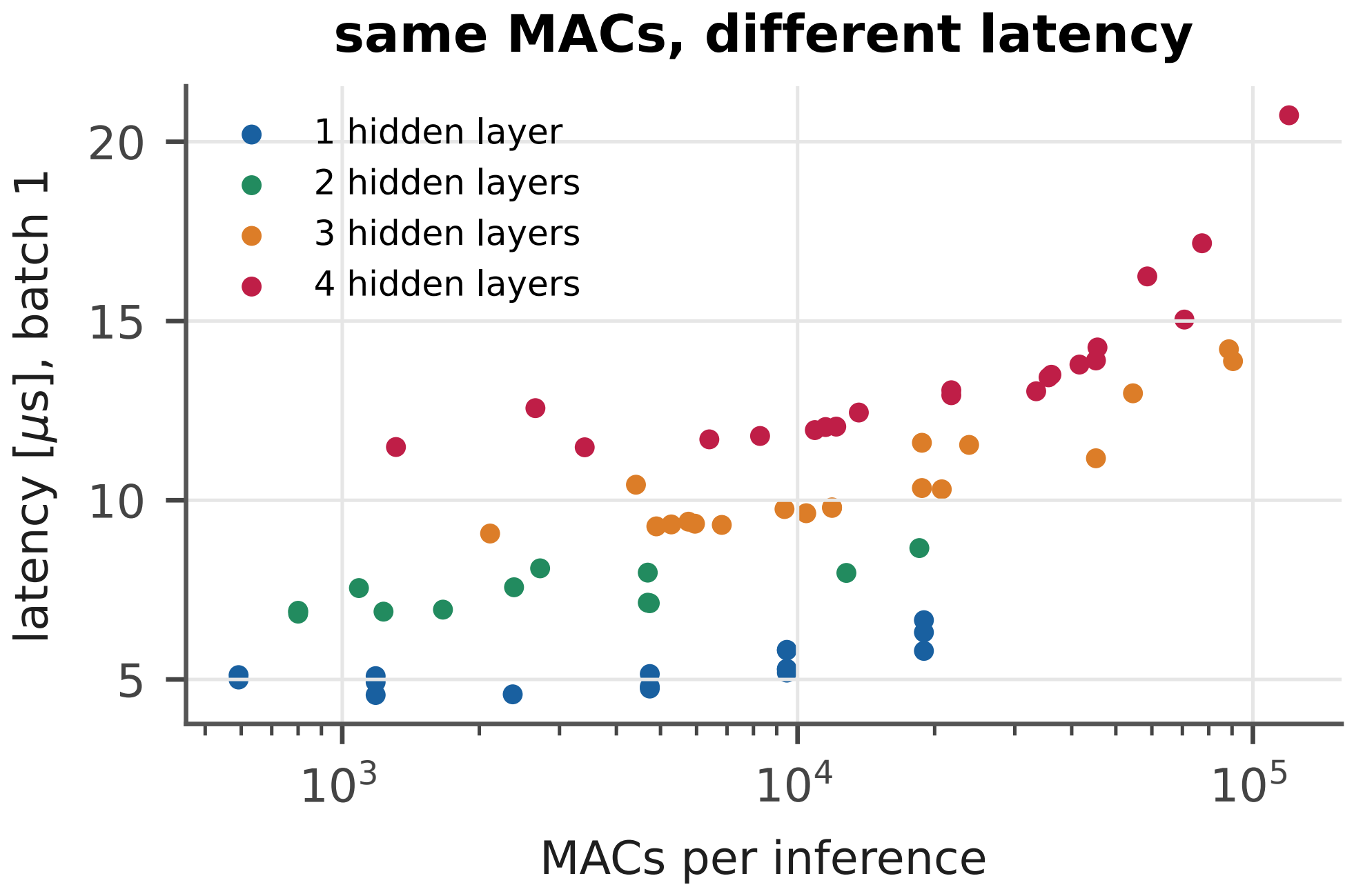

same MACs, different latency
1 hidden layer
2 hidden layers
3 hidden layers
4 hidden layers
latency [μs], batch 1
20
15
10
5
10^3
10^4
10^5
MACs per inference