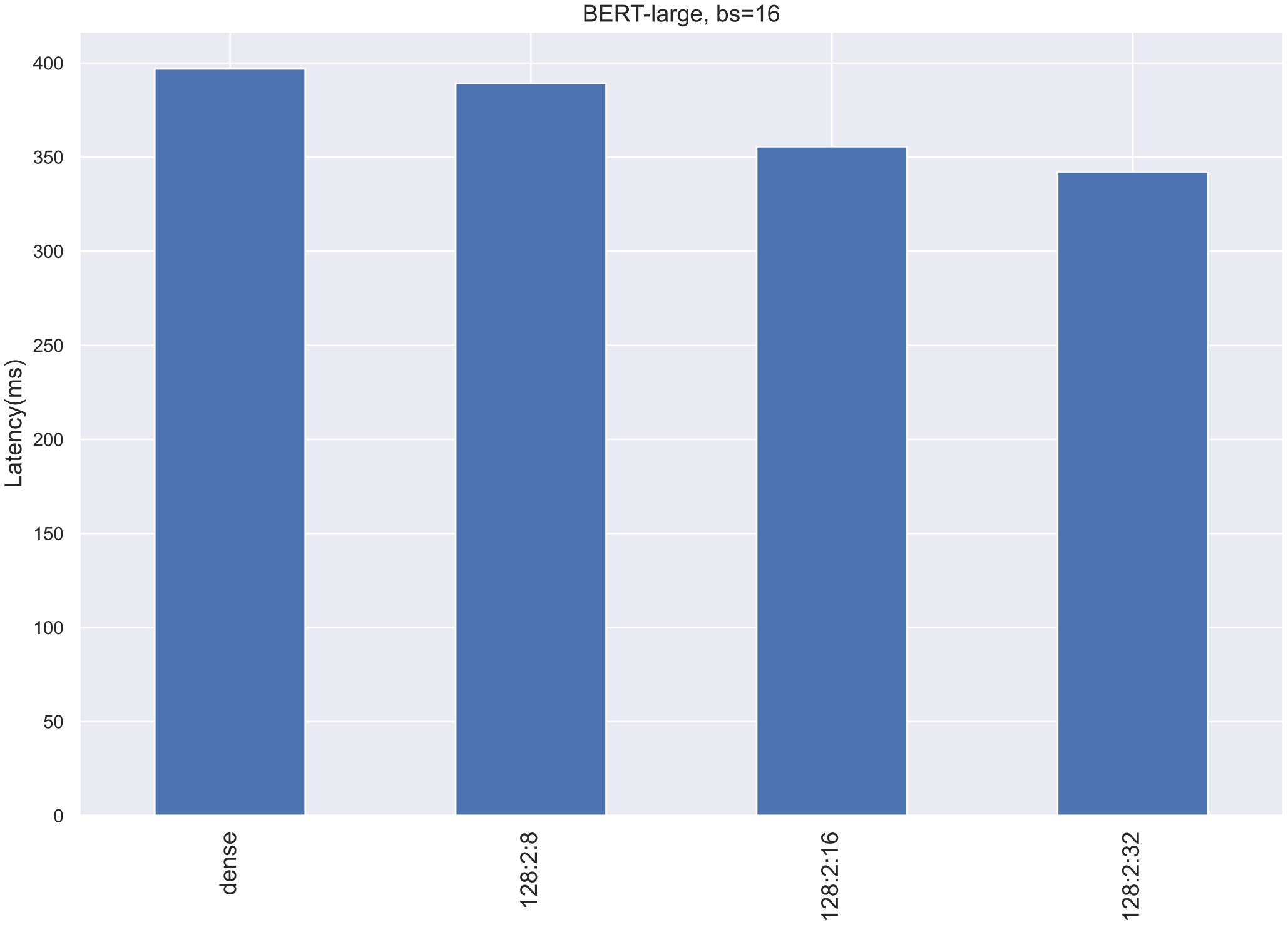
BERT-large, bs=16
Latency(ms)
400
350
300
250
200
150
100
50
0
dense
128:2:8
128:2:16
128:2:32