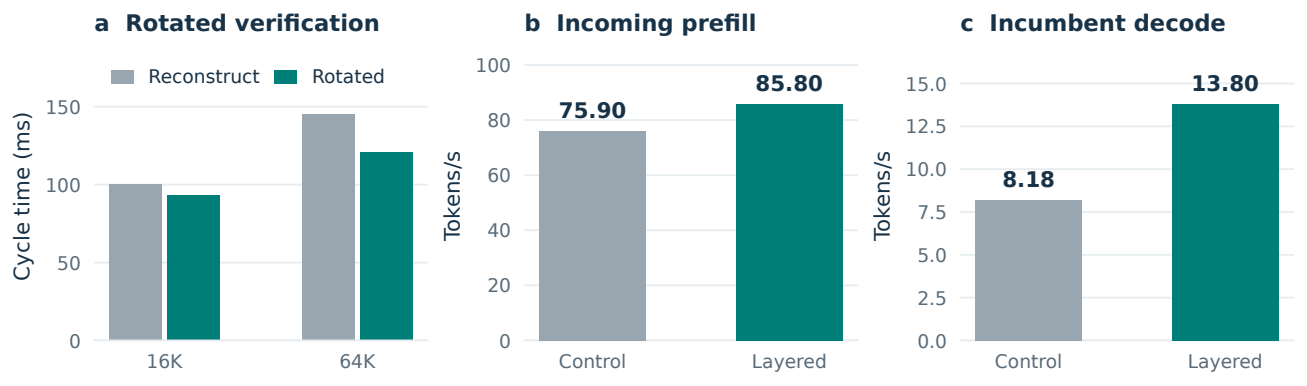

a Rotated verification
Reconstruct
Rotated
Cycle time (ms)
150
100
50
0
16K
64K
b Incoming prefill
Tokens/s
100
80
60
40
20
0
75.90
85.80
Control
Layered
c Incumbent decode
Tokens/s
15.0
12.5
10.0
7.5
5.0
2.5
0.0
8.18
13.80
Control
Layered