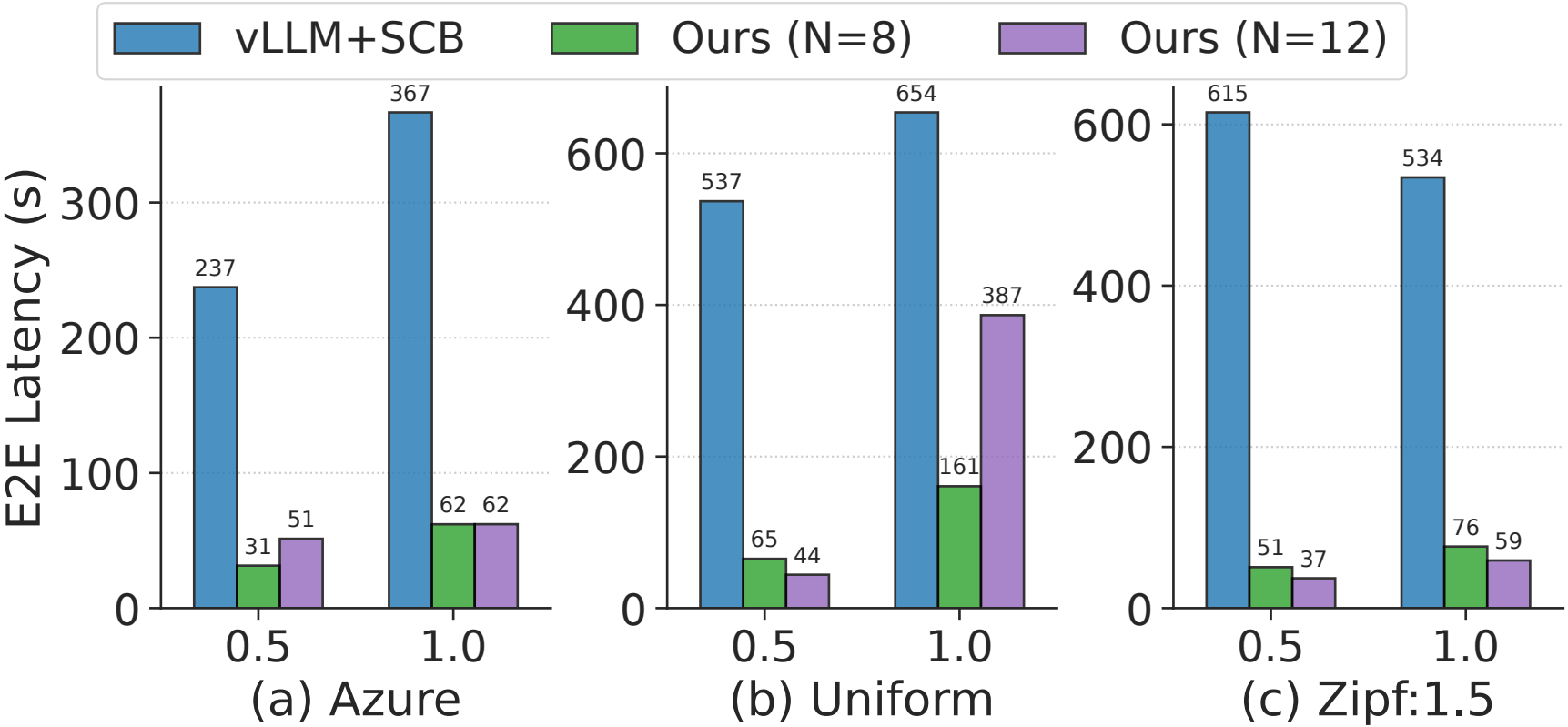
vLLM+SCB
Ours (N=8)
Ours (N=12)
E2E Latency (s)
0
100
200
300
237
31
51
367
62
62
0.5
1.0
(a) Azure
0
200
400
600
537
65
44
654
161
387
0.5
1.0
(b) Uniform
0
200
400
600
615
51
37
534
76
59
0.5
1.0
(c) Zipf:1.5