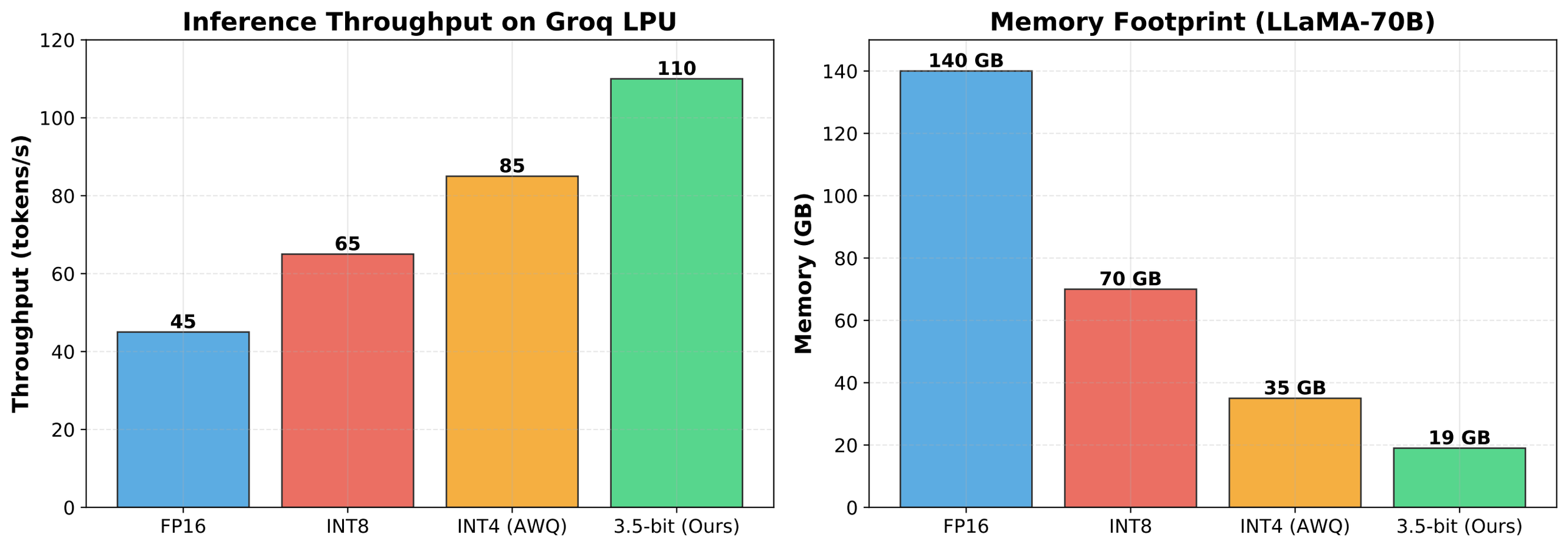
Inference Throughput on Groq LPU
Throughput (tokens/s)
0
20
40
60
80
100
120
45
65
85
110
FP16
INT8
INT4 (AWQ)
3.5-bit (Ours)
Memory Footprint (LLaMA-70B)
Memory (GB)
0
20
40
60
80
100
120
140
140 GB
70 GB
35 GB
19 GB
FP16
INT8
INT4 (AWQ)
3.5-bit (Ours)