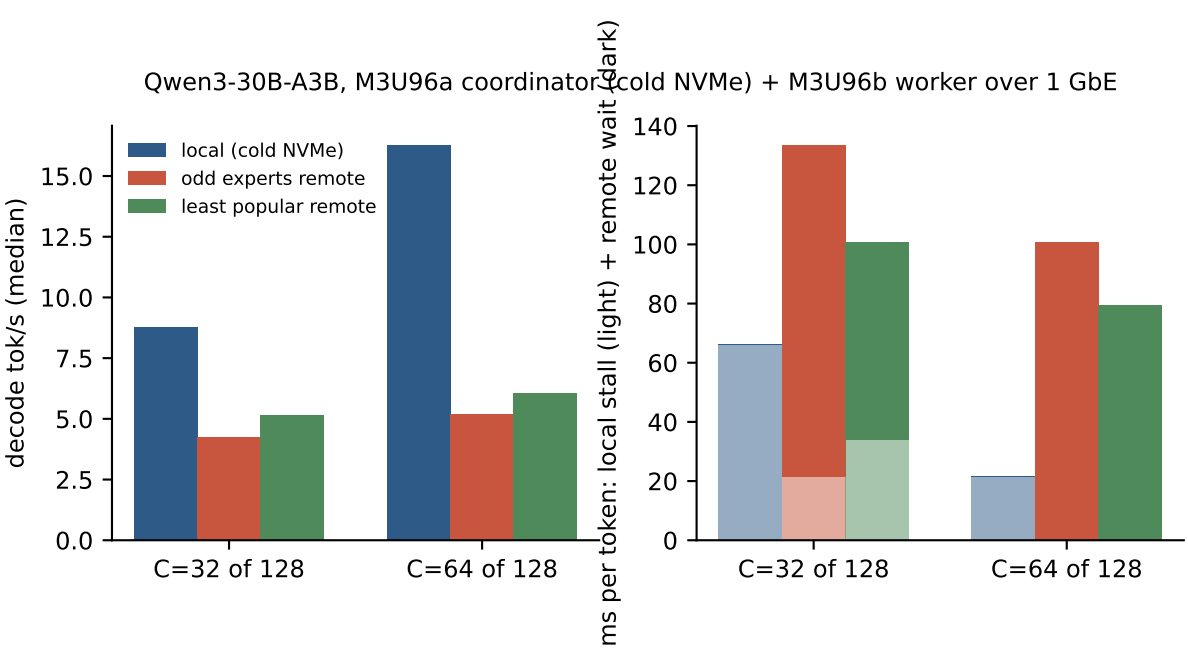

Qwen3-30B-A3B, M3U96a coordinator (cold NVMe) + M3U96b worker over 1 GbE
local (cold NVMe)
odd experts remote
least popular remote
decode tok/s (median)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
C=32 of 128
C=64 of 128
ms per token: local stall (light) + remote wait (dark)
0
20
40
60
80
100
120
140
C=32 of 128
C=64 of 128