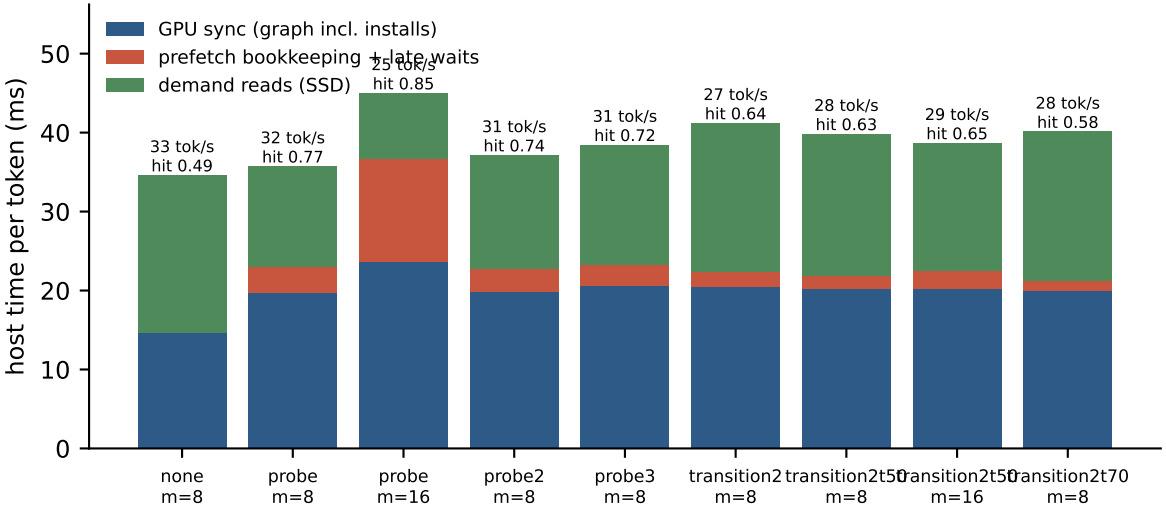

OLMoE, 16 slots, cold NVMe (M5 Max): where the decode step waits
GPU sync (graph incl. installs)
prefetch bookkeeping + late waits
demand reads (SSD)
host time per token (ms)
0
10
20
30
40
50
33 tok/s
hit 0.49
32 tok/s
hit 0.77
25 tok/s
hit 0.85
31 tok/s
hit 0.74
31 tok/s
hit 0.72
27 tok/s
hit 0.64
28 tok/s
hit 0.63
29 tok/s
hit 0.65
28 tok/s
hit 0.58
none
m=8
probe
m=8
probe
m=16
probe2
m=8
probe3
m=8
transition2
m=8
transition2t50
m=8
transition2t50
m=16
transition2t70
m=8