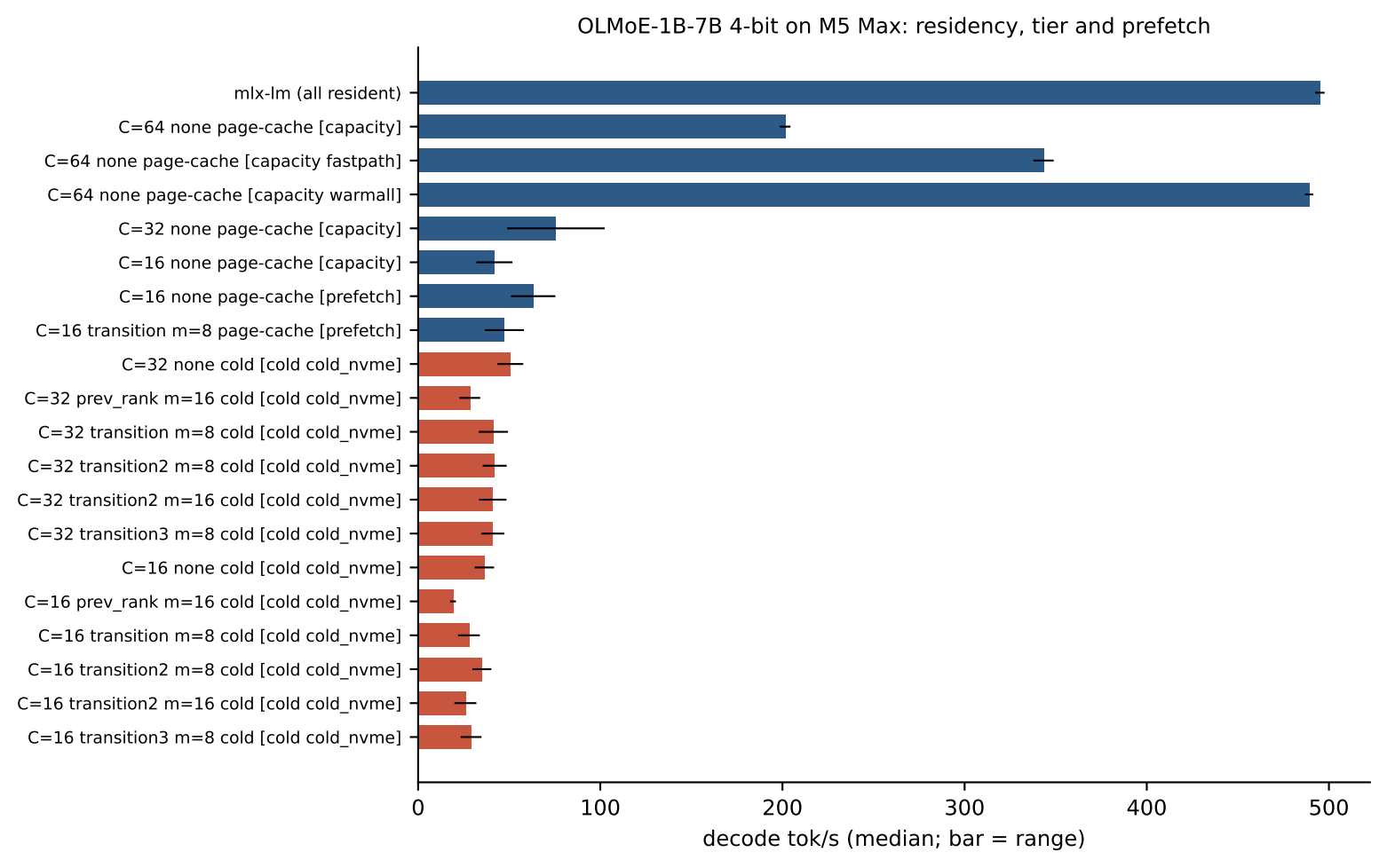

OLMoE-1B-7B 4-bit on M5 Max: residency, tier and prefetch
mlx-lm (all resident)
C=64 none page-cache [capacity]
C=64 none page-cache [capacity fastpath]
C=64 none page-cache [capacity warmall]
C=32 none page-cache [capacity]
C=16 none page-cache [capacity]
C=16 none page-cache [prefetch]
C=16 transition m=8 page-cache [prefetch]
C=32 none cold [cold cold_nvme]
C=32 prev_rank m=16 cold [cold cold_nvme]
C=32 transition m=8 cold [cold cold_nvme]
C=32 transition2 m=8 cold [cold cold_nvme]
C=32 transition2 m=16 cold [cold cold_nvme]
C=32 transition3 m=8 cold [cold cold_nvme]
C=16 none cold [cold cold_nvme]
C=16 prev_rank m=16 cold [cold cold_nvme]
C=16 transition m=8 cold [cold cold_nvme]
C=16 transition2 m=8 cold [cold cold_nvme]
C=16 transition2 m=16 cold [cold cold_nvme]
C=16 transition3 m=8 cold [cold cold_nvme]
0
100
200
300
400
500
decode tok/s (median; bar = range)