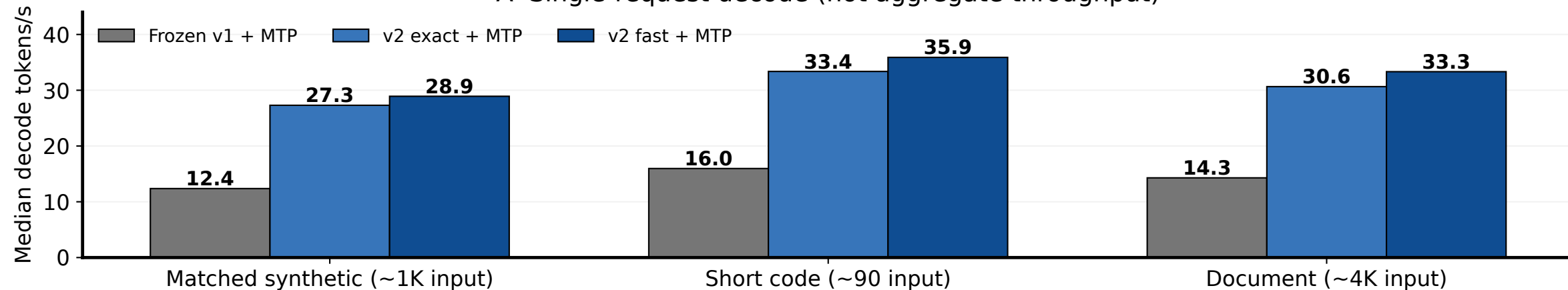


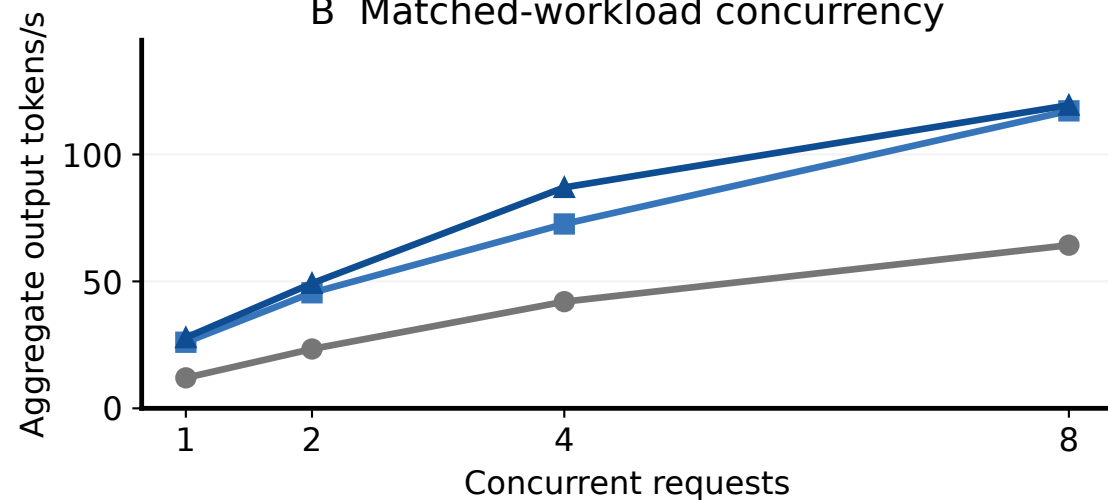
Runtime v2 — faster serving, unchanged NVFP4 weights

Final validation on one DGX Spark • W4A4 + MTP • October 31, 2026

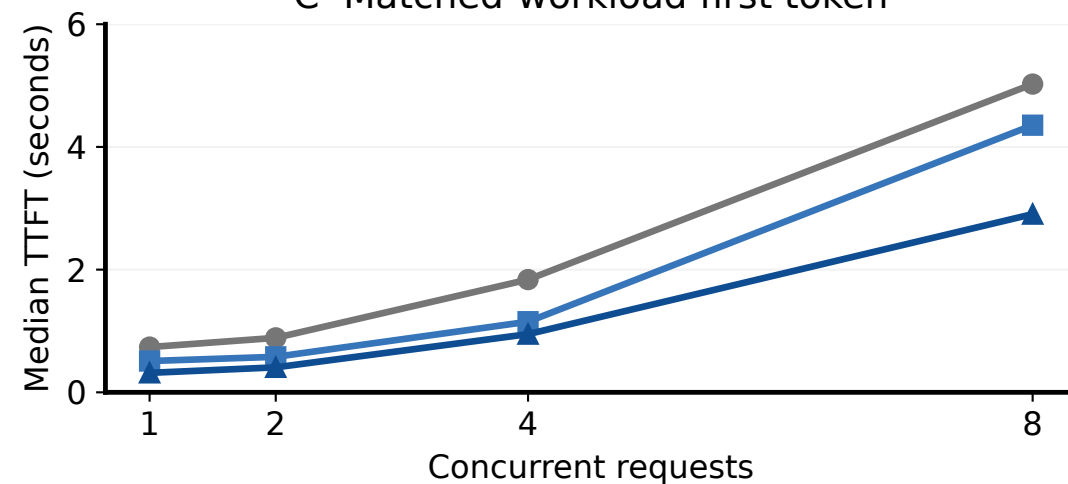
A Single-request decode (not aggregate throughput)



B Matched-workload concurrency



C Matched-workload first token



Exactly 256 output tokens • greedy • no prefix-cache hits • 0 errors in final performance runs

Matched: 8 requests/level; code/document: 4 at c=1 • final results, not earlier screening peaks

Exact preserves reference routing math; fast changes FP32 reduction order and may change generated text.

Source: runtime-v2/runs/final_*/{perf_matched,screen_code,screen_doc}.json