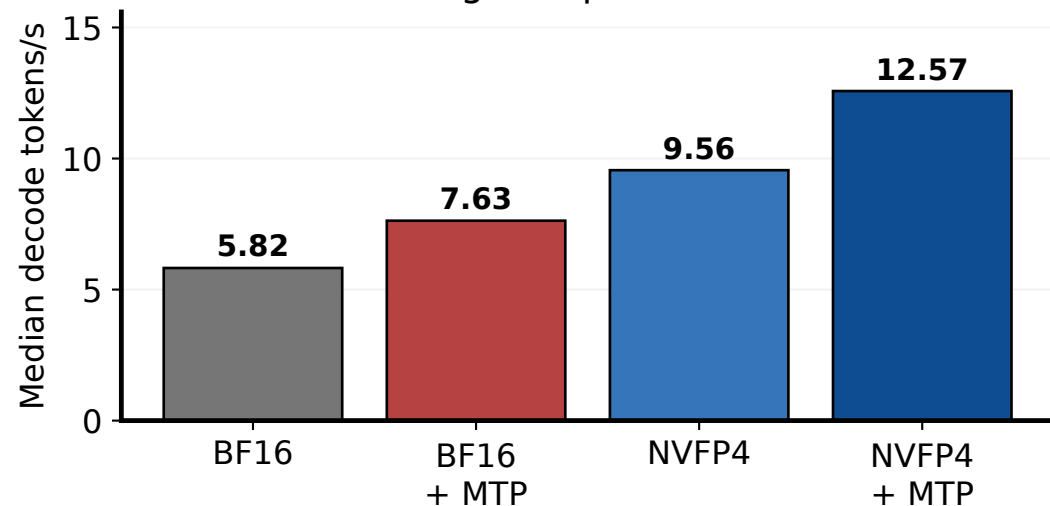


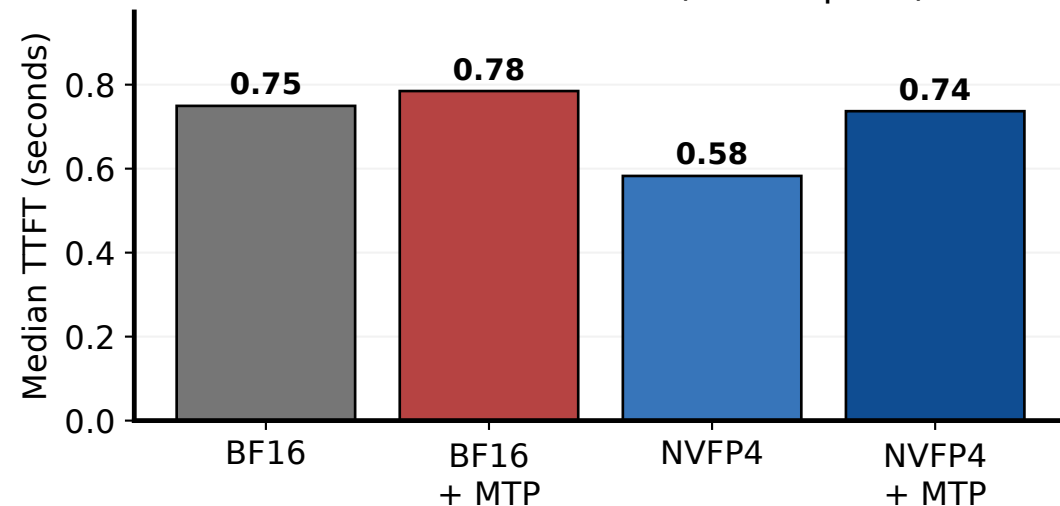
SenseNova Loop 9B NVFP4 — one DGX Spark

Functional release baseline • native W4A4 • original weights, PTQ only

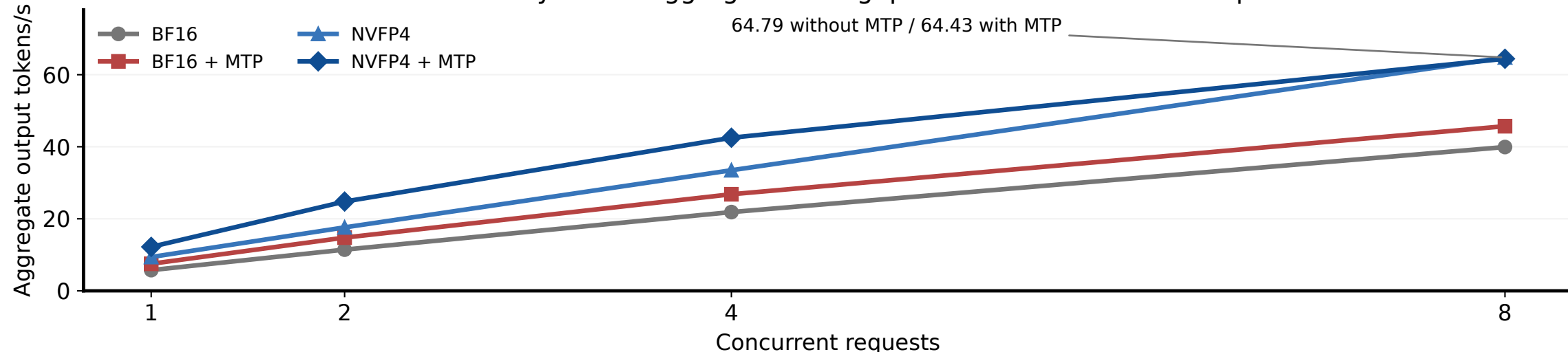
A Single-request decode



B Time to first token (one request)



C Concurrency scales aggregate throughput — not individual-user speed



1,051 prompt tokens → 256 output tokens • greedy • unique prompts, no prefix-cache hits

16 requests/level without MTP; 8 with MTP • CUDA graphs • 0 request errors/OOMs

Source: evaluation/perf_sglang_*.json • Measurements: October 30, 2026