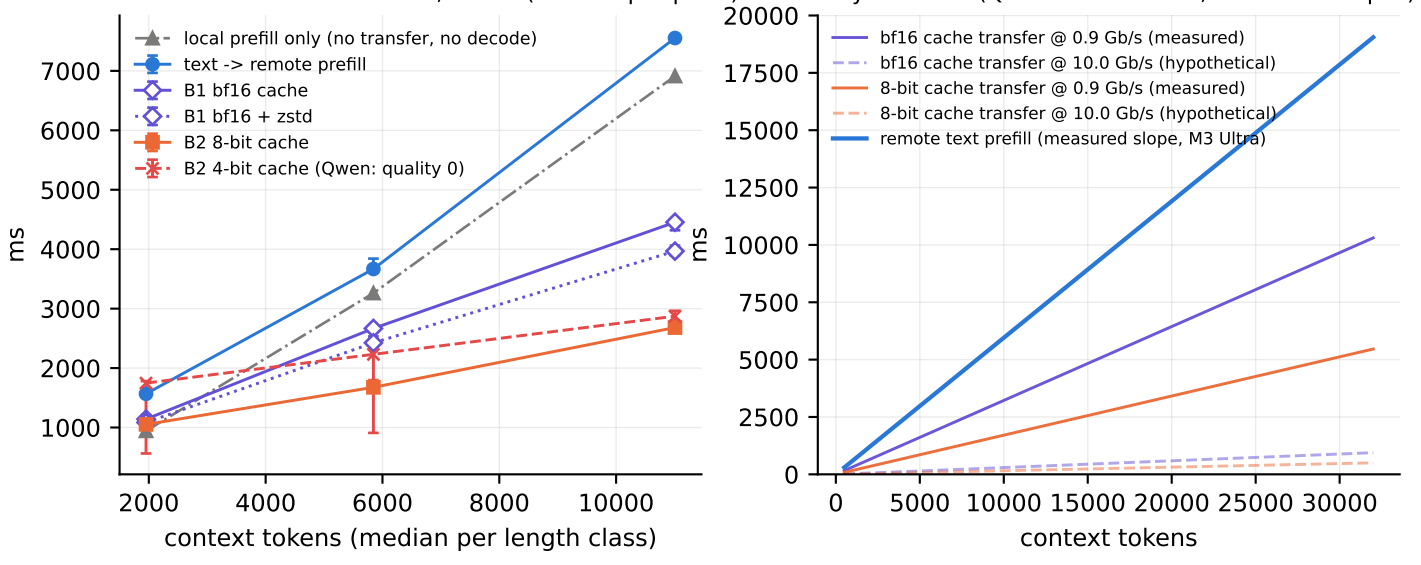

Measured: M3U96a -> M3U96b, 1 GbE (18 runs per point)
local prefill only (no transfer, no decode)
text -> remote prefill
B1 bf16 cache
B1 bf16 + zstd
B2 8-bit cache
B2 4-bit cache (Qwen: quality 0)
ms
context tokens (median per length class)
Analytic model (Qwen2.5-3B cache, measured slopes)
bf16 cache transfer @ 0.9 Gb/s (measured)
bf16 cache transfer @ 10.0 Gb/s (hypothetical)
8-bit cache transfer @ 0.9 Gb/s (measured)
8-bit cache transfer @ 10.0 Gb/s (hypothetical)
remote text prefill (measured slope, M3 Ultra)
ms
context tokens