PURETENSOR
INTELLIGENCE DISTILLED

Speculative Decoding Wins the Benchmark and Loses the Seat

Two models, two engines, one trade: single-stream speed for KV cache and concurrency

Paper: PT-R-2026-019 v1.0

Author: PureTensor Ltd

Date: 12 September 2026

Status: Published

Abstract

Speculative decoding proposes several tokens from a small draft model (or a bundled draft head), verifies them in one forward pass of the full model, and keeps the longest accepted prefix. When acceptance is high, each step yields multiple tokens and single-stream throughput rises. We tested this technique on two production seats in our sovereign fleet: a mixture-of-experts model served on two workstation Blackwell GPUs, and a 475 GiB mixture-of-experts model spanning two nodes over a 200 GbE RDMA fabric. Both checkpoints ship their own multi-token-prediction heads; neither seat had used them.

On the resident seat, enabling the draft head raised single-stream output from 108.3 tokens per second to 174.6, a factor of 1.61. This exceeded our researched expectation of about 1.2. The cost was a 17.5 per cent reduction in KV-cache capacity, a 14 per cent drop in aggregate throughput at 64 concurrent requests (26 per cent against the warm control), a 3.7-fold increase in cold time-to-first-token on long prompts, and a prefix-cache hit rate that fell from 0.56 to 0.45. On the large seat the exchange was starker: single-stream rose from 27–28 tokens per second to 95–178 (3.5 times on the greedy set), but the KV pool shrank from roughly 890,000 tokens to 260,352, the concurrency ceiling dropped from at least 32 slots (the launch cap, still scaling) to 8, and the engine hung three times during draft-on runs against zero hangs in about 60 minutes of draft-off serving.

Neither seat shipped the change. Both serve bursty, multi-consumer traffic where aggregate throughput, pool size, and cold latency determine the user experience. The single-stream benchmark that community reports quote is exactly the metric a shared seat cares least about. The patched images, the gate batteries, and the raw logs are retained for future work, including a planned test of the engine's option to disable speculation above a configurable concurrency threshold.

1. Motivation and Prior Expectation

Speculative decoding is attractive because it promises higher throughput without changing the model's weights or its final distribution. A draft head, a small extra layer trained alongside the main model and shipped in the same checkpoint, generates candidate continuations cheaply; the full model scores them in parallel and accepts the longest prefix whose tokens match what the full model would have sampled. When acceptance is high, one forward pass emits several

tokens, and the wall-clock cost per token falls.

Community benchmarks and upstream issue-tracker reports on our GPU class suggested a single-stream gain of 10 to 35 per cent, with about 20 per cent likely. The same sources warned that aggregate throughput at high concurrency could be flat or as much as 20 per cent lower, because the draft head consumes memory that would otherwise hold additional KV-cache entries. We set a ship rule before the first run: aggregate at 64 concurrent requests must reach at least 95 per cent of baseline, cold time-to-first-token must not regress beyond a stated bound, and the prefix-cache hit rate must hold. The experiment would pass or fail on those gates, not on single-stream speed.

2. Experiment A: The Resident Seat

The resident seat serves every agent on the fleet. It runs Qwen3.8-Flash-Next in the GPU vendor's NVFP4 quantisation (4-bit routed experts), served by vLLM at tensor-parallel 2 on two workstation Blackwell GPUs (96 GB each, PCIe, no NVLink). The context window is 262,144 tokens, the sequence limit is 256, and GPU memory utilisation is set to 90 per cent. Traffic is bursty and multi-consumer.

The checkpoint ships a multi-token-prediction head. The seat had never enabled it. Our first two attempts to load the draft head failed with the same error: a draft-layer expert parameter had no destination in the weight map. A research sweep of the engine source and its upstream history revealed two separate defects. First, the draft-layer builder remaps the quantisation exclusion lists to the draft layer's runtime name but never remaps the per-layer method map, so the lookup misses. Second, the mixed-precision configuration's routed-experts branch dispatches only FP8, NVFP4, W4A16-NVFP4, and MXFP8, then returns nothing; a block-scaled FP8 draft layer therefore falls through to an unquantised method. Patching either defect alone reproduces the identical error.

A fix had been merged upstream on 8 September 2026, two days before we encountered the failure. It could not be cherry-picked because the upstream module had been renamed. We hand-transposed six hunks across two files into our pinned image. A separate trap delayed the work: the engine reads its quantisation map from the model's configuration file, not from the separate quantisation-config file we had been editing, so every metadata change we tried earlier had no effect.

With the patch applied, the draft head loaded. The engine log confirmed the expected path: the FP8 block scales of the draft layer were refined from [128, 128] to [64, 64] to fit the tensor-parallel-sharded intermediate size of 320, and a Triton kernel was selected. No expert-parallel flag was required; the global flag would have converted all 48 main-model NVFP4 layers to expert parallelism over PCIe for no kernel benefit.

Metric	Baseline image, no MTP	New image, MTP off (control)	New image, MTP k=2
Single-stream tok/s (20-prompt greedy, 512 max)	108.3	104.0 cold / 100.8 warm	174.6 (1.61×)
Aggregate tok/s at 64 concurrent, 256 generated	3,196	2,967 cold / 3,751 warm	2,764 (0.86× baseline)
GPU KV cache, tokens	3,580,560	3,580,560	2,953,102 (0.825×)
Max concurrency at full 262k context	13.65×	13.65×	11.27×
Cold TTFT, ~12k-token prompt	0.66 s	not run	2.46 s (3.7×)

Metric	Baseline image, no MTP	New image, MTP off (control)	New image, MTP k=2
Warm TTFT / prefix-cache hit rate	0.10 s / 0.56	0.10 s / 0.56	0.19 s / 0.45
Draft acceptance / mean accepted length	n/a	n/a	0.817 / 2.63
Greedy byte-match vs baseline, 20 prompts	n/a	20/20 cold, 15/20 warm	5/20

The 1.61-fold single-stream gain exceeded our researched expectation. Draft acceptance was 0.817; of 1,870 draft positions, 1,642 were accepted at position 0 and 1,413 at position 1, yielding a mean accepted length of 2.63 tokens per step. Tool calls and JSON-schema outputs parsed correctly.

The run failed every ship gate. Aggregate throughput at 64 concurrent was 2,764 tokens per second, 14 per cent below baseline and 26 per cent below the warm control. Cold time-to-first-token on a roughly 12,000-token prompt rose from 0.66 seconds to 2.46 seconds. KV-cache capacity fell by 17.5 per cent, and the prefix-cache hit rate dropped from 0.56 to 0.45. We restored the seat to the baseline image. Downtime for the two loading attempts and the A/B comparison was about 26 minutes per phase, including 190–220 seconds of engine loads.

3. Experiment B: The Large Seat

The large seat runs DeepSeek-V4.1-Flash, a 475 GiB mixture-of-experts checkpoint with MXFP4 routed experts, served by SGLang at tensor-parallel 4 and expert-parallel 4 across two nodes (two workstation Blackwell GPUs each) over a 200 GbE RDMA fabric. The context window is 409,600 tokens. The checkpoint ships a bundled draft head that adds about 2 GB per rank.

With the draft enabled, the concurrency ceiling is 8 slots at 85 per cent memory fraction. Attempts to boot with 16 or 32 slots fail because the draft's sliding-window pool leaves no room for the full KV pool. The KV pool with the draft holds 260,352 tokens; without the draft it holds roughly 890,000 tokens.

Configuration	Single-stream tok/s	Aggregate at 8	Aggregate at 16	Aggregate at 32
Draft off (k=1), 512-token input	27–28	312	509	777
Draft off (k=1), 2,048-token input	27–28	not recorded	526	813
Draft on, 512-token input, 8,192 out	178.3	628.3	cannot boot	cannot boot
Draft on, 8,192-token input, 8,192 out	151.8	642.3	cannot boot	cannot boot

On the 20-prompt greedy set, draft-on delivered 95 tokens per second single-stream against 27–28 with the draft off, a factor of 3.5. Draft acceptance ranged from 0.76 to 0.91 across runs; the captured run showed acceptance of 0.81 and a mean accepted length of 5.04 tokens.

Without the draft, single-stream is bound by per-step latency (cross-node synchronisation plus a host-side callback that fetches embedding rows), not by compute. Batching is therefore nearly free: per-request rate rises from 27–28 tokens per second single-stream to 39 tokens per second per request at 8 concurrent, and aggregate was still scaling at 32 concurrent, the launch cap. The draft's advantage narrows as concurrency rises.

Greedy outputs with the draft on matched outputs with the draft off on 7 of 20 prompts. But draft-off versus a draft-off rerun matched only 6 of 20, and draft-on versus draft-on rerun

likewise 6 of 20. The engine's greedy decode is documented as not bitwise stable across batch composition, so the draft adds no measurable divergence beyond the engine's own non-determinism. Divergences were stylistic; arithmetic and factual checks held.

Stability was worse with the draft. Three scheduler-watchdog hangs occurred across the draft-on runs; zero hangs occurred in about 60 minutes of draft-off serving, including a full 94-item run of our internal frontier benchmark (agentic 92.1, scorer 96.0, flat 93.5, 100 per cent coverage). The hang was later traced to a host-side row-fetch stall that the draft's variable gather sizes provoke; that forensic is the subject of a separate paper.

The large seat continues to run with the draft off. Its consumers are batch fan-out and long agent runs where aggregate throughput and pool size matter more than single-stream speed.

4. The Shared Shape

Both experiments exhibit the same trade-off: speculation converts KV-cache memory and concurrency headroom into single-stream speed. On the resident seat, 1.61 times single-stream cost 17.5 per cent of the KV pool, 14–26 per cent of aggregate throughput, and 3.7 times cold time-to-first-token. On the large seat, 3.5 times single-stream cost about 70 per cent of the KV pool and three-quarters of the concurrency ceiling, plus observable stability regressions.

Neither seat is interactive-single-user. Both serve bursty, multi-consumer traffic where aggregate throughput, pool size, and cold latency determine the experience. The single-stream benchmark that community reports quote is exactly the metric a shared seat cares least about. A benchmark that measures only single-stream speed will show speculation as a clear win; a benchmark that measures aggregate throughput, memory headroom, and tail latency under concurrent load will show it as a loss on these workloads.

5. Harness Lessons

Three observations from the resident-seat runs bear on future experiments.

First, live-seat greedy outputs are only comparable with a cold prefix cache. Warm replays reuse KV blocks computed under other batch shapes and flip near-tie tokens. The control image matched baseline 20 of 20 cold but only 15 of 20 warm; the five mismatches were exactly the prompts the cache had kept. Any byte-identity gate should follow an engine restart.

Second, single-stream throughput varied from 104.0 to 100.8 tokens per second between two identical runs with zero consumer traffic and idle GPUs. Treat plus or minus 5 per cent single-stream as noise on this seat.

Third, the prefix-cache hit counter on a live seat includes the cold pass, so a fivefold replay reads about 0.56 rather than 0.8. Relative gates are more informative than absolute thresholds.

6. What These Numbers Will Not Carry

The sample sizes are small. Each configuration was measured once; the 20-prompt greedy set is the only repeated workload. Single-stream noise of plus or minus 5 per cent means the 1.61-fold gain carries about plus or minus 5 per cent of uncertainty, and the 3.5-fold gain on the large seat has similar uncertainty. The large-seat matrix is one pass per cell.

The resident-seat comparison is confounded by the image change. The control image (new image, MTP off) was 4 per cent slower single-stream than baseline cold and 7 per cent slower warm, even though the only change was the six-hunk patch. The aggregate drop under MTP is 14 per cent against baseline but 26 per cent against the warm control. We do not know how much of the aggregate loss is due to the draft and how much to the patch or to measurement

variance.

The large-seat engine is documented as not bitwise stable across batch composition, so the 7-of-20 greedy match between draft-on and draft-off is not meaningfully worse than the 6-of-20 match between two draft-off runs. The divergence metric does not distinguish draft-induced error from engine-induced non-determinism.

The three scheduler-watchdog hangs on the large seat occurred during draft-on runs, but the root cause (a host-side row-fetch stall) was identified only after the measurement window closed. We cannot rule out that the hangs were triggered by the specific prompts or batch shapes rather than by the draft head itself.

The untested next step on the resident seat (k=1 with prefix caching disabled, and the engine's disable-speculation-above-N-running-requests option) might recover the aggregate loss while preserving the single-stream gain for interactive traffic. We have not run it.

Claims that survive: speculation raises single-stream speed on both seats (1.61 times and 3.5 times on the greedy sets), at the cost of KV-cache capacity (17.5 per cent and roughly 70 per cent), aggregate throughput at high concurrency (14–26 per cent on the resident seat), cold time-to-first-token (3.7 times on the resident seat), and, on the large seat, concurrency ceiling (from 32 to 8 slots) and observed stability. Claims that do not survive: any precise estimate of the aggregate loss attributable solely to the draft head, any claim about output quality beyond stylistic divergence, and any prediction about the hybrid option that disables speculation above a concurrency threshold.

7. Next Steps

We retain the patched images and the gate batteries for both seats. The next experiment on the resident seat will test the engine's disable-speculation-above-N-running-requests option, which would enable the 1.61-fold gain for the interactive minority while avoiding the aggregate loss at high concurrency. A separate paper will report the forensic on the large-seat scheduler hang and the host-side row-fetch stall that the draft's variable gather sizes provoke.

PureTensor operates a sovereign AI infrastructure fleet (on-premises GPU compute, storage, and serving) run day-to-day by autonomous agents under human direction. Measurements were taken on 10 and 11 September 2026; raw artefacts are retained. This paper is PT-R-2026-019.