FP8 and Per-Tenant Speculative Decoding from 8B to 32B

FP8 with per-tenant EAGLE3 draft heads served 2.40×, 1.866× and 2.051× the on-time requests per GPU of BF16 at 8B, 14B and 32B

Pranav Varshney
Nodus Compute

Viswa Kotra
Nodus Compute

Abstract

Interactive serving capacity is determined by how many requests finish within their latency objective, not raw token throughput. We evaluate one serving configuration across Qwen3-8B, 14B, and 32B on real GPUs: FP8 weights with an off-the-shelf EAGLE3 draft head activated per tenant by a research serving harness built for this study, a configuration not yet in the product. Against BF16 serving, the combined configuration reaches 18.5333 versus 7.72 SLO-met requests per GPU-second on 8B (2.40×), 30.52 versus 16.35 on 14B (1.866×), and 15.95 versus 7.775 on 32B (2.051×), with 0 serving errors in every arm. FP8 alone reaches 17.1067 on 8B, 21.85 on 14B, and 11.04 on 32B. On the 8B bursty 16-tenant trace, output tokens per second remain nearly constant while TTFT p95 falls from 6.332 s to 0.0366 s, showing that the gain is work moved inside the SLO rather than extra generated tokens. Quality is evaluated with GSM8K and MMLU using paired bootstrap confidence intervals. The 8B combined arm passes the preregistered point-estimate gate, although the GSM8K interval extends to −1.21 points; FP8 alone fails that gate. At 14B and 32B every measured quality-delta interval includes 0. The components are standard and each larger model uses one seed, so the contribution is a unified systems measurement and cross-scale characterization, not a new decoding algorithm.

1 Introduction

A saturated decoder can report high token throughput while bursty requests wait in a queue and miss their latency target. Those late tokens consume the same GPU time but do not satisfy an interactive service commitment. We therefore use SLO-met requests per GPU-second as the primary serving-capacity metric and retain output throughput, TTFT, TPOT, errors, and task accuracy as guard measurements.

FP8 reduces the bandwidth and capacity cost of model weights [1]. Speculative decoding uses a smaller draft path to propose tokens that the target model verifies in parallel [2]. Neither technique is new. The systems question is whether a multi-tenant serving layer can activate the pair per tenant, produce a stable gain under offered load, and keep quality within a declared gate across model sizes.

This paper unifies the 8B and scale-out campaigns, and it absorbs report 015, FP8 and Speculation at 14B and 32B, whose measurements are the section on scaling to Qwen3-14B and Qwen3-32B and the quality results for both models. The 8B experiment uses a fixed bursty 16-tenant trace and exposes the queueing mechanism in detail. The 14B and 32B

experiments use a load ladder to find each arm's best SLO-qualified point. Together they show which part of the gain comes from FP8, what speculation adds, how the ratio changes with model size, and where the quality evidence is strong or incomplete.

2 Experimental design

Hardware and models. Every arm uses one GPU for its model. The campaign covers Qwen3-8B, Qwen3-14B, and Qwen3-32B [3]. The larger-model runs specify vLLM v0.30.0 [4].

Arms. Arm A is BF16 serving. Arm B changes weights to FP8. Arm C uses FP8 with a publicly released EAGLE3 draft head [5] activated by the research serving harness built for this study. This three-arm design separates the contribution of quantization from the incremental contribution of speculative verification.

8B traffic. All arms replay the same 5,560-request bursty trace from 16 tenants for 300 s. Arrivals, prompts, and output lengths are identical. The throughput gate requires Arm C to reach at least 1.5× Arm A.

14B and 32B traffic. Each arm runs a 120 s offered-load ladder with at most 256 output tokens per request. A request meets SLO when TTFT is at most 1 s and TPOT at most 0.1 s. We report the best SLO-met requests per GPU-second over the ladder. Each model uses one seed.

Quality. We evaluate GSM8K and MMLU. The 8B campaign uses all 1,319 GSM8K problems and a seeded 2,000-question MMLU subset under greedy decoding. Deltas use 10,000 paired bootstrap resamples and 95% confidence intervals [6]. The preregistered 8B quality gate is based on the point estimate being at least −1.0 points relative to BF16. The larger-model campaign requires its declared quality gates and reports whether each interval includes 0.

3 Qwen3-8B under bursty multi-tenant traffic

Table 1: Measured 8B serving results, one GPU, 5,560 requests, 300 s per arm. Latencies are seconds.

Arm	SLO req/GPU-s	Out tok/s	TTFT p50 / p95	Errors
A: BF16	7.72	4,719.4	1.6196 / 6.332	0
B: FP8	17.1067	4,728.7	0.0315 / 1.8184	0
C: FP8 + draft	18.5333	4,738.4	0.0254 / 0.0366	0

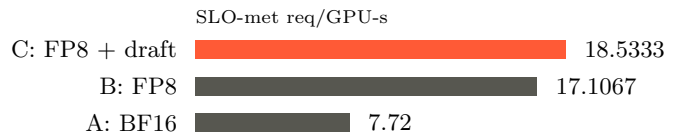


Figure 1: SLO-met requests per GPU-second on Qwen3-8B.

Arm C delivers 2.40× Arm A and clears the 1.5× gate. Yet output token throughput differs by less than 0.5% across arms. The capacity gain is therefore not a large change in generated tokens per second; it is a queueing result in which the same aggregate work reaches request deadlines.

TPOT p50/p95 falls from 0.0079/0.0084 s in BF16 to 0.0057/0.0065 s in FP8 and 0.0028/0.0038 s in the combined arm. FP8 removes most queueing but retains a TTFT p95 of 1.8184 s. The draft head collapses that remaining tail to 0.0366 s. All arms record 0 errors.

4 Scaling to Qwen3-14B and Qwen3-32B

Table 2: Best SLO-qualified point on Qwen3-14B, one GPU, one seed.

Arm	SLO req/GPU-s	Attainment	Ratio to BF16
A: BF16	16.35	1.0	baseline
B: FP8	21.85	1.0	1.336×
C: FP8 + EAGLE3	30.52	1.0	1.866×

Table 3: Best SLO-qualified point on Qwen3-32B, one GPU, one seed.

Arm	SLO req/GPU-s	Attainment	Ratio to BF16
A: BF16	7.775	0.945	baseline
B: FP8	11.04	1.0	1.420×
C: FP8 + EAGLE3	15.95	0.9995	2.051×

The combined arm clears the 1.5× gate at both larger sizes. FP8 alone contributes 1.336× at 14B and 1.420× at 32B. Adding EAGLE3 raises the total to 1.866× and 2.051×. On 32B, the best BF16 rung has only 0.945 attainment while the combined arm holds 0.9995, so the ratio includes both higher admitted rate and a larger fraction delivered inside SLO.

The cross-model result is not monotonic in parameter count: the 8B ratio is 2.40×, the 14B ratio 1.866×, and the 32B ratio 2.051×. Draft acceptance, memory behavior, the load ladder, and the BF16 knee all affect the ratio. The correct generalization is that every tested size clears the same 1.5× threshold, not that one constant multiplier applies to all models.

5 Quality evaluation

Table 4: 8B accuracy. The combined arm passes the point-estimate gate; FP8 alone does not.

Arm	GSM8K	MMLU	Delta interpretation
A: BF16	92.646	63.45	reference
B: FP8	91.433	63.40	GSM8K −1.21; fails
C: FP8 + draft	92.494	63.45	−0.15 / 0.00; passes

For the 8B combined arm, GSM8K changes by −0.15 points with a 95% CI of −1.21 to +0.91, and MMLU changes by 0.00 with a CI of −0.80 to +0.75. The point estimates clear the preregistered −1.0 gate. The GSM8K interval extends below −1.0, so the data do not exclude a regression of that magnitude. FP8 alone loses 1.21 GSM8K points with a CI of −2.35 to −0.08 and fails.

At 14B and 32B, every reported quality-delta interval includes 0 and every preregistered gate passes. The evidence is still one seed per model and two benchmark families. It supports unchanged measured benchmark quality, not universal semantic equivalence.

6 Component attribution and operating point

The three-arm design shows that FP8 is necessary but not sufficient for the full result. It delivers most of the 8B gain and roughly one third to two fifths improvement at larger sizes. The draft head then reduces sequential target-model work and, at 8B, removes the remaining queueing tail. The combined result depends on both the per-token execution path and the offered-load operating point.

A raw tokens-per-second benchmark would miss the 8B result because all three arms generate about 4.7k output tokens per second. SLO-met requests per GPU-second captures the actual capacity boundary: the point where arrivals, queueing, TTFT, and TPOT combine into completed interactive work.

Table 5: Larger-model accuracy. Every paired-bootstrap delta interval against BF16 includes 0.

Model	Arm	GSM8K	MMLU
32B	BF16 / FP8 / combined	94.92 / 95.15 / 94.77	77.80 / 77.25 / 77.25
14B	BF16 / FP8 / combined	95.15 / 94.62 / 94.54	74.95 / 74.95 / 74.95

6.1 Evidence boundary

The study covers three Qwen3 sizes, one GPU per model, one bursty trace at 8B, a load ladder at 14B and 32B, and one seed for each larger model. EAGLE3 heads are off-the-shelf and the paper does not claim a new algorithm. The 8B quality interval contains a possible regression larger than the point-estimate gate. Results should not be transferred to other models, draft heads, prompts, hardware, or SLO definitions without a matched measurement.

7 Reproducibility

A reproduction must pin the model revision, precision, serving engine, draft head, GPU type, prompt set, output cap, arrivals, and SLO thresholds. Each arm must receive identical requests, be charged for the same window and drain policy, and report errors. The load-ladder studies must expose all rungs so the selected best point can be checked for bracketing rather than chosen from an unreported search.

Quality evaluation must reuse the same item ordering and greedy decoding, then compute paired—not independent—bootstrap intervals over questions. The 8B full GSM8K set and seeded 2,000-question MMLU subset should remain distinct from the larger-model evaluation rather than being pooled.

8 Conclusion

Across Qwen3-8B, 14B, and 32B on one GPU per model, FP8 with a per-tenant off-the-shelf speculative draft head served 1.866–2.40× as many requests inside SLO as BF16, with 0 serving errors. The 8B trace shows that the gain comes from latency and queueing rather than a large change in raw token throughput. Quality passes the declared point-estimate gate at 8B and has intervals including 0 at 14B and 32B, with the 8B GSM8K interval retained as an uncertainty. The result is a reproducible serving configuration, measured in research and not yet in the product, and a cross-scale measurement: standard components integrated at the tenant boundary, evaluated on the outcome interactive systems are paid to deliver.

References

[1] P. Micikevicius et al. FP8 Formats for Deep Learning. arXiv preprint, 2022.

[2] Y. Leviathan, M. Kalman, Y. Matias. Fast Inference from Transformers via Speculative Decoding. ICML 2023.

[3] Qwen Team. Qwen3 Technical Report. arXiv preprint, 2025.

[4] W. Kwon et al. Efficient Memory Management for Large Language Model Serving with PagedAttention. SOSP 2023.

[5] Y. Li et al. EAGLE-3: Scaling up Inference Acceleration of Large Language Models via Training-Time Test. arXiv preprint, 2025.

[6] B. Efron, R. J. Tibshirani. An Introduction to the Bootstrap. Chapman and Hall, 1993.