PURETENSOR
INTELLIGENCE DISTILLED

Fastest Is Not Best

The one-node seat beats the two-node record on half the hardware

Paper: PT-R-2026-021 v1.0

Author: PureTensor Ltd

Date: 18 September 2026

Status: Published

Abstract

A GPU vendor released, on 9 September 2026, an NVFP4 quantisation of GLM-5.3-Flash, a mixture-of-experts model. The checkpoint totals 204.5 GB across 44 files. Our fleet's two GPU workstations each carry two Blackwell GPUs with 205.3 GB raw VRAM, joined by a 200 GbE RDMA fabric; the quantisation therefore cannot fit on one node even at 100 per cent utilisation with zero KV cache. We benchmarked it cross-node at tensor-parallel 4 and found aggregate throughput of 117.9 tokens per second at single concurrency, rising to 763.9 at 16 concurrent streams. An initial write-up described this as "the fastest seat ever measured on this fleet". That claim was false: it had been compared only with earlier cross-node runs, never with the incumbent one-node seat.

Re-benchmarking the same day against the live one-node seat (Qwen3.8-Flash-Next, a different mixture-of-experts model served at tensor-parallel 2) showed that the one-node configuration beats the two-node configuration on every metric except single-stream decode. At 16 concurrent streams the one-node seat delivers 896.9–982.5 aggregate tokens per second against 763.9 for the cross-node run. Prefill throughput at approximately 31,000 tokens of context is 7,412 tokens per second on one node, 4,711 on two. The one-node seat also scores higher on our internal frontier benchmark (94 original items, ten dimensions, every verdict produced by code, no language-model judge): 96.2 agentic and 96.8 flat, against 88.2 and 90.8 for the one-node EXL3 quantisation of GLM-5.3-Flash.

The cross-node NVFP4 result is an improvement over an earlier NVFP4 variant of the same model: plus 18 per cent at single concurrency, plus 11 per cent at 16 concurrent. It is the best GLM-5.3-Flash we have measured. It is still slower than a different model running on half the hardware. The paper documents three measurement traps encountered during the work (a prefix-cache hit mistaken for a prefill measurement, an EXL3 seat whose throughput swings threefold across warm-up passes, and a response-field mismatch that made a working model look broken) and two open defects (an intermittent RDMA local-protection fault, a grammar-engine failure on the EXL3 seat). We adopt a rule: before benchmarking, check whether the checkpoint bytes exceed the raw VRAM of a single node; a number that can only be bought by evicting seven services across both nodes is a lab result, not a seat.

1. The Fleet and the Question

Our fleet comprises two GPU workstations. Each carries two workstation-class Blackwell GPUs, each GPU offering 97,887 MiB (together 205.3 GB raw per node). The nodes are joined by a 200 GbE RDMA fabric. A "seat" is a model served continuously for the fleet's autonomous agents; a seat that requires both nodes evicts every other GPU service on both.

On 9 September 2026 the GPU vendor published an NVFP4 quantisation of GLM-5.3-Flash. The checkpoint spans 44 files and totals 204.5 GB. The quantisation keeps attention, the router, shared experts and the vision tower in BF16; only the routed experts are 4-bit group-16; KV cache is FP8; no multi-token-prediction shards are included. The obvious question was whether this checkpoint could become the fleet's primary seat.

2. Experimental Setup

We benchmarked the vendor NVFP4 quantisation cross-node at tensor-parallel 4. The harness, container image and deviation thresholds were unchanged from our 2 September cross-node run of an earlier NVFP4 variant of the same model. All runs used 512-token answers, 131,000 tokens of context, FP8 KV cache, 75 per cent utilisation and multi-token prediction disabled. Concurrency levels tested were 1, 4, 8 and 16 concurrent streams.

After the cross-node run we re-benchmarked, on the same day with identical scripts, against the live one-node seat (Qwen3.8-Flash-Next, served at tensor-parallel 2 on one node) and the one-node EXL3 quantisation of GLM-5.3-Flash.

Sample sizes were as follows: one cross-node sweep per concurrency level; two passes for the Qwen seat at 16 concurrent (both values reported); three passes for the EXL3 seat (ranges reported). Benchmark scores for the internal frontier benchmark came from earlier runs on the same placements.

3. Measured Results

Table 1 presents the core throughput and latency figures.

Metric	Qwen3.8-Flash-Next, 1 node	GLM-5.3-Flash EXL3, 1 node	GLM-5.3-Flash NVFP4, 2 nodes
Aggregate tok/s, 1 concurrent	108.5	91.5 (79–104 across passes)	117.9
Aggregate tok/s, 4 concurrent	347.6	115.2 (74–116)	341.8
Aggregate tok/s, 8 concurrent	573.7	328.9 (118–339)	543.8
Aggregate tok/s, 16 concurrent	896.9 / 982.5	234.0 (207–303)	763.9
Prefill tok/s at ~31k prompt	7,412	3,650	4,711
TTFT p50 / p95	35 / 805 ms	110 / 111 ms	not recorded
End-to-end p50, 64-token turn	606 ms	660 ms	not recorded
Per-stream decode tok/s, 1 concurrent	108.6	112.8	124.2
Internal frontier benchmark, agentic / flat	96.2 / 96.8	88.2 / 90.8	not run

The cross-node NVFP4 run showed per-stream decode rates falling as concurrency rose: 124.2 at single concurrency, 92.7 at 4, 74.1 at 8, 51.5 at 16. Time to first token followed the inverse

pattern: 226 ms at single concurrency, 466 at 4, 626 at 8, 796 at 16. Engine initialisation took 71.6 seconds cold and 17.0 seconds warm.

Against the earlier NVFP4 variant measured on 2 September, the new quantisation gained 18 per cent at single concurrency, 16 per cent at 4, 14 per cent at 8, and 11 per cent at 16. Prefill improved by 4 per cent. This is the best GLM-5.3-Flash we have measured on this fleet.

4. The Comparison That Was Not Made

The first write-up called the cross-node result "the fastest seat ever measured on this fleet". That statement was false. The comparison had been drawn only against the cross-node tensor-parallel-4 history, never against the incumbent one-node seat.

Table 1 shows the correction. Qwen3.8-Flash-Next on one node beats the vendor GLM quantisation on two nodes on every metric except single-stream decode. At 16 concurrent streams the one-node seat delivers 896.9–982.5 aggregate tokens per second; the two-node configuration delivers 763.9. Prefill throughput is 7,412 tokens per second on one node against 4,711 on two. Time to first token at p50 is 35 ms on one node; the cross-node run did not record TTFT but its single-concurrency figure of 226 ms is already seven times slower.

The one-node seat also beats the one-node GLM EXL3 quantisation on every metric except per-stream decode at single concurrency (112.8 versus 108.6 tokens per second). These are different models, so throughput alone does not establish which seat is more useful. The internal frontier benchmark provides a capability comparison: Qwen3.8-Flash-Next scores 96.2 agentic and 96.8 flat; GLM-5.3-Flash EXL3 scores 88.2 and 90.8. The throughput advantage and the capability advantage point the same way.

5. Why the Two-Node Number Cannot Become a Seat

The vendor NVFP4 checkpoint is 204.5 GB. One node is 205.3 GB raw. The checkpoint cannot run on one node even at 100 per cent utilisation with zero KV cache. Any practical serving configuration requires headroom for the KV cache, so the margin is illusory.

Every NVFP4 variant of this model is in the same position. All of them keep attention in BF16, and the three community variants are 197.9, 199.4 and 187.7 GB. None fits on one node with usable KV headroom.

The only quantisations that fit are the EXL3 family, which quantise attention as well as the experts. The 4-bit-per-weight variant is 175.8 GB; the 3.51-bit-per-weight variant is 156.7 GB; a 2-bit-class variant is 97.7 GB. These fit, but as Table 1 shows, the EXL3 seat is slower than the Qwen seat at every concurrency level above 1.

A cross-node seat evicts every other GPU service on both nodes. In our fleet that means evicting the resident agent model on one node and six services on the other: a 27B model, a small reasoning model, two document-processing replicas, the memory embedder and an OCR model. Seven evictions is not a seat; it is a lab result.

We now apply a rule before benchmarking: compare raw VRAM against checkpoint bytes. If the checkpoint exceeds a single node's capacity, any performance number is a ceiling that cannot be realised without evicting the fleet's working services.

6. Three Measurement Traps

Three errors arose during the benchmarking and were caught before publication.

The first involved a repeated prefill measurement that returned 95,941 tokens per second. This was a prefix-cache hit, not a prefill measurement. Only a fresh prompt measures prefill. The 4,711 figure in Table 1 comes from fresh prompts.

The second involved the EXL3 seat's warm-up behaviour. Its five-minute warm-up script assumes the seat is ready after that interval. In practice, single-stream throughput climbed from 79.4 to 91.5 to 104.0 tokens per second across three consecutive sweeps. Aggregate throughput at 8 and 16 concurrent streams swung threefold between passes. The cause is that the EXL3 engine's adaptive multi-token prediction re-tunes depth per batch bucket, and prefix caching is enabled while the harness sends identical prompts. One EXL3 pass is not a measurement; we report medians across passes and note the ranges. By contrast, the Qwen seat's passes repeated to within 1 per cent.

The third involved a response-field mismatch. The cross-node build returns its reasoning tokens in a field named "reasoning", not the field the client expected. When the token budget is consumed by reasoning, the expected content field is empty. This looks like a broken model. It is not; the model is working and the client is looking in the wrong place.

7. Two Open Defects

Two defects remain unresolved.

The first is an intermittent RDMA local-protection fault. The first 16-concurrent run killed the engine. One node's fabric adapter reported a queue-pair error; the other reported a remote-access error; the engine logged a remote process exit. The fault did not reproduce; the subsequent 16-concurrent runs completed cleanly. This is a fabric fault, not a model fault. It is also distinct from a multi-token-prediction defect we have seen on this model (that defect manifests as a sampling RPC timeout when speculation is enabled; this run had speculation disabled). We have not yet tried the available mitigations: disabling DMA-buffer registration or lowering the GPU-direct level.

The second is a grammar-engine failure on the one-node EXL3 seat. Plain chat requests with no tools attached trigger an error on the model's special-token IDs. This defect must be resolved before that seat carries tool work.

8. Operational Notes

Both container images had been pruned from the nodes and required re-pulling: 42.2 GB for one image, 29 GB for the other. The efficient procedure is to pull once on one node and stream the image across the fabric. A second internet pull competes with any download in flight and stalls both.

Engine initialisation times differed between configurations. The cross-node NVFP4 seat took 71.6 seconds cold and 17.0 seconds warm. The one-node seats were not timed separately, but their warm-up behaviour (stable for Qwen, unstable for EXL3) is documented above.

9. What These Numbers Will Not Carry

The sample sizes are small. The cross-node configuration was measured in a single sweep per concurrency level. The Qwen seat at 16 concurrent was measured in two passes; we report both values (896.9 and 982.5) rather than a mean, because two observations do not yield a reliable mean. The EXL3 seat was measured in three passes; we report ranges. The internal frontier benchmark scores come from earlier runs on the same placements, not from runs performed on the measurement day.

The comparison is between different models. Qwen3.8-Flash-Next and GLM-5.3-Flash are both mixture-of-experts models, but they differ in architecture, training data and capability profile. Throughput comparisons across different models are confounded by capability differences. We address this by including the internal frontier benchmark scores, which show the Qwen seat scoring higher on both agentic and flat tasks (96.2 / 96.8 versus 88.2 / 90.8). The benchmark was not run on the cross-node NVFP4 configuration; we assume it would match or closely approach the one-node EXL3 scores, since it is the same model at higher precision, but this assumption is untested.

The TTFT and end-to-end latency figures for the cross-node configuration were not recorded in the same format as for the one-node configurations. The per-concurrency TTFT figures (226 / 466 / 626 / 796 ms) are available, but the p50 and p95 breakdown is not. Direct latency comparisons are therefore incomplete.

The intermittent RDMA fault did not reproduce, so we cannot confirm whether it would recur under sustained load. The grammar-engine failure on the EXL3 seat affects only tool-bearing requests, but we have not characterised its frequency or trigger conditions.

The rule we adopt (checkpoint bytes versus raw VRAM, checked before benchmarking) is a heuristic. It does not account for KV cache requirements, activation memory or other runtime overheads. A checkpoint that fits by this rule may still fail to serve at useful context lengths.

10. What Changes

We do not change the fleet's primary seat. Qwen3.8-Flash-Next remains the incumbent. The cross-node GLM-5.3-Flash configuration is the best we have measured for that model, but it is slower than the incumbent on half the hardware and would evict seven services to run.

The next experiment is to resolve the grammar-engine failure on the one-node EXL3 seat and re-benchmark it with tools enabled. If that seat can match the Qwen seat's throughput on tool-bearing workloads, it becomes a candidate for tasks where GLM-5.3-Flash's capability profile is preferred. We will also test the available mitigations for the RDMA fault, though the fault's failure to reproduce makes controlled testing difficult.

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 September 2026; raw artefacts are retained. This paper is PT-R-2026-021.