

Local Ollama GPU Performance Consolidation

Captured 2026-09-10T06:22:31.530890+00:00. Hardware: RTX 3080 Ti (12 GB) + GTX 1080 Ti (11 GB), NVIDIA driver 580.178.04, PHB topology without NVLink.

Scope. All retained local benchmark JSON artifacts plus completed session measurements from the RTX 3080 Ti + GTX 1080 Ti evaluation. The companion JSON preserves full raw source artifacts and provenance hashes. Baseline snapshot: Included.

Why these tests were run

Stage	Why	What was tested	Decision supported
1. Hardware and topology discovery	The RTX 3080 Ti and GTX 1080 Ti have different generations, capacities, and no NVLink; a single-model split could therefore be slower than independent workers.	NVIDIA visibility, VRAM, driver state, and independent vs split Qwen3 8B placement.	Use GPU 0 for latency-sensitive Qwen work. Treat GPU 1 as independent capacity, not automatically pooled VRAM.
2. Qwen3 8B production baseline	Qwen3 8B is the existing archive model, so its direct 8k and 40k behavior set the production routing baseline.	GPU 0 alone, GPU 1 alone, independent concurrent requests, and a 40k two-GPU split.	Independent requests increased aggregate throughput; a 40k split was substantially slower. Qwen3 8B also cannot meet Hermes 64k context because it caps at 40,960 tokens.
3. Large-model candidate comparison	Laguna XS Q4 and Qwen3-Coder 30B were candidates for a 64k Hermes session.	Both-GPU 8k/40k throughput, then RTX-only host-RAM-mapped baselines and both-GPU measurements at 8k, 40k, and 64k.	Laguna was fastest at 64k; Qwen3-Coder remained viable when coding specialization matters. The RTX-only baseline proved the tradeoff between freeing GPU 1 and generation speed.
4. Multi-GPU placement and KV-cache experiments	The question was whether Ollama automatically prioritizes the faster RTX 3080 Ti as context grows.	Automatic layer placement, a manual 3:2 RTX-favoring tensor split, F16/F16 versus Q8/Q8 KV cache, and an incomplete K=Q8/V=F16 run.	Ollama auto-placement primarily fit weights and KV cache; it did not preferentially place more frequently used layers on the RTX. Manual 3:2 splits OOMed. Q8/Q8 improved ingestion but reduced generation; the mixed test is deliberately recorded as incomplete.
5. Locally installed Qwen3.8 27B evaluation	The already-downloaded local Qwen candidate needed evidence before considering further downloads.	Qwen3.8:latest at 8k, 40k, and 64k across both GPUs.	It accepted 64k but generation degraded from 21.4 tok/s at 8k to 1.74 tok/s at 64k, so it is not the practical 64k Hermes choice on this hardware.
6. Conservative production restoration	The user requested more archive work but correctly questioned whether a 50% increase was justified by the uneven GPU performance.	A dedicated local GPU 1 Ollama service and disjoint email-worker partitions, followed by a scheduled successful GPU 1 batch and a short throughput sample.	Use an 80% GPU 0 / 20% GPU 1 shard split, not a 50% work claim. The first 8.4-minute parallel sample was faster but remains preliminary; the immutable baseline snapshot supports a later comparison.

Interpretation limits

- Token rates use controlled prompts, not a full distribution of archive-message lengths.
- The initial parallel email sample is too short to establish a final completion estimate.
- Q8 KV-cache tradeoffs need task-quality validation, not only throughput comparison.
- No claim is made for intentionally stopped, failed, or source-artifact-overwritten experiments.

Validated findings

- For qwen3:8b, independent per-GPU jobs are faster in aggregate than splitting one 40k request across both GPUs.
- qwen3:8b has an effective 40,960-token cap and is not a 64k Hermes model.
- Laguna XS Q4 was the fastest tested 64k text-agent model on this machine.
- Qwen3-Coder 30B is viable at 64k but slower than Laguna under the measured layouts.
- Qwen3.8 27B supports 64k but generation fell sharply as context grew on this hardware.
- Q8/Q8 KV cache improved 64k prompt ingestion but reduced generation speed versus the F16/F16 default for the tested large models.

Benchmark measurements

Model	Context	Layout	KV cache	Prompt tok/s	Generation tok/s	Wall s	Evidence
laguna-xs-2.1:q4_K_M	8k	RTX 3080 Ti only; host-RAM mapped	F16/F16 default	314.2	50.3	71.7	recorded isolated-layout artifact
laguna-xs-2.1:q4_K_M	8k	Unspecified	F16/F16 default	1,850.8	97.5	31.5	recorded benchmark artifact
laguna-xs-2.1:q4_K_M	40k	RTX 3080 Ti only; host-RAM mapped	F16/F16 default	478.7	45.3	90.8	recorded isolated-layout artifact
laguna-xs-2.1:q4_K_M	40k	Unspecified	F16/F16 default	874.7	62.1	60.4	recorded benchmark artifact
laguna-xs-2.1:q4_K_M	64k	Both GPUs	F16/F16 default	629.8	48.5	—	session-measured result preserved in report
laguna-xs-2.1:q4_K_M	64k	Both GPUs	Q8/Q8	760.0	45.3	92.9	recorded isolated-layout artifact
laguna-xs-2.1:q4_K_M	64k	RTX 3080 Ti only; host-RAM mapped	F16/F16 default	448.4	41.1	136.6	recorded isolated-layout artifact
qwen3-coder:30b	8k	Unspecified	F16/F16 default	1,742.1	66.3	57.6	recorded benchmark artifact
qwen3-coder:30b	40k	Unspecified	F16/F16 default	449.7	42.3	104.6	recorded benchmark artifact
qwen3-coder:30b	64k	Both GPUs	F16/F16 default	364.1	23.1	186.7	recorded isolated-layout artifact
qwen3-coder:30b	64k	Both GPUs	Q8/Q8	364.1	23.1	—	session-measured result preserved in report
qwen3.8:latest	8k	Both GPUs	F16/F16 default	405.7	21.4	42.6	recorded isolated-layout artifact
qwen3.8:latest	40k	Both GPUs	F16/F16 default	265.1	4.0	195.0	recorded isolated-layout artifact
qwen3.8:latest	64k	Both GPUs	F16/F16 default	214.2	1.7	357.3	recorded isolated-layout artifact
qwen3:8b	8k	GTX 1080 Ti only	F16/F16 default	726.5	42.2	16.3	recorded benchmark artifact
qwen3:8b	8k	GTX 1080 Ti only	F16/F16 default	722.9	42.1	16.5	recorded benchmark artifact
qwen3:8b	8k	RTX 3080 Ti only	F16/F16 default	4,412.7	110.7	5.9	recorded benchmark artifact
qwen3:8b	8k	RTX 3080 Ti only	F16/F16 default	4,417.7	109.6	6.1	recorded benchmark artifact
qwen3:8b	40k	Both GPUs	F16/F16 default	683.1	39.8	63.2	recorded benchmark artifact
qwen3:8b	40k	RTX 3080 Ti only	F16/F16 default	2,629.7	68.1	19.2	recorded benchmark artifact
qwen3:8b	64k	RTX 3080 Ti only; host-RAM mapped	F16/F16 default	3,361.0	85.5	13.7	recorded isolated-layout artifact

Actual long-context acceptance

Model	Result
qwen3:8b	Requested 64k was capped at 40,960 tokens.
laguna-xs-2.1:q4_K_M	Accepted a real 57,043-token prompt at requested 65,536 context.
qwen3-coder:30b	Accepted a real 57,009-token prompt at requested 65,536 context.
qwen3.8:latest	Accepted a real 57,012-token prompt at requested 65,536 context.

Current conservative batch configuration

Qwen3 8B email analysis uses an 8k context: GPU 0 owns shards 0–3 of 5, and GPU 1 owns shard 4. The split is disjoint and intended to reduce end-to-end completion time conservatively rather than force a 50% throughput claim.

Provenance

The complete machine-readable report is `/home/ai/files/google-lewismoten/meta/performance-reports/20260910-local-ollama-gpu-performance-consolidated.json`. The following retained source files were embedded into that JSON:

Artifact	SHA-256
gpu_ollama_benchmark_results.json	f064f9c10a54cf4e2399d6a175d1230fcb71e4595f174191a8695bf35dc576d8
laguna-xs-2.1_q4_K_M_gpu_benchmark_results.json	23ef43a6af819a95da353c22709c71315d7cfaf76a868061d75e0d749e3d45a0
qwen3-coder_30b_gpu_benchmark_results.json	be973e2777126af991a1974bb5a79d97825f94d7d406864db7c98929444f719c
laguna-xs-2.1_q4_K_M_gpu0_only_8k_one.json	ffb489a06f650c3d666b63aee13bf25f9b2589d4074825b03de4ece4450ac541
laguna-xs-2.1_q4_K_M_gpu0_only_40k_one.json	7c0cb775a5078c7c4c46da60414c56f08697ef9c9772d2250de88d35069beb0f
laguna-xs-2.1_q4_K_M_gpu0_only_64k_one.json	3c43c7c631d86fbec777eae9f91b889a646ca62f46b8328ca23221e313457b99
laguna-xs-2.1_q4_K_M_both_gpus_64k_one.json	d159bf5c6364857c661271d295df7f0776a3e181e1286ffe3f5435a0f7a28bb1
qwen3-coder_30b_both_gpus_64k_one.json	610a45267596dabcf8bb54f2b53d38554313a98ebdb52512b13ab70f6a19877f
qwen3_8b_gpu0_only_64k_one.json	893a599e3466575ed636ac4143ee191087552707bd12acd0c3458a4b5cedbd2d
qwen3.8_latest_both_gpus_8k_k-f16_v-f16_one.json	c53bab6b78589efcd823520868a55a9a132d57939f354c18436ab695084c9db7
qwen3.8_latest_both_gpus_40k_k-f16_v-f16_one.json	22d60c767c745a957d13f0db24a7dfd9a1875d2f19bd1faa8cdd05ba4a5f17ee
qwen3.8_latest_both_gpus_64k_k-f16_v-f16_one.json	d05cf396a1423fd791942e2c774daf1ea8633fd43e275dbe461d1746cdba46f5

Incomplete or failed experiments

- Laguna 64k K=Q8/V=F16 did not complete; it is not reported as a metric.
- Manual 3:2 tensor splits at 64k F16 KV exhausted GPU 0 memory for both Laguna and Qwen3-Coder.