

PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

AIGENESIS — PCIe x1 vs x16: Does It Matter for Local AI Inference?

Agentic AI Systems | 24 October 2026 | Technical benchmark publication | Document ID: 202610-AASPUB-0009

1. Research question and test configuration

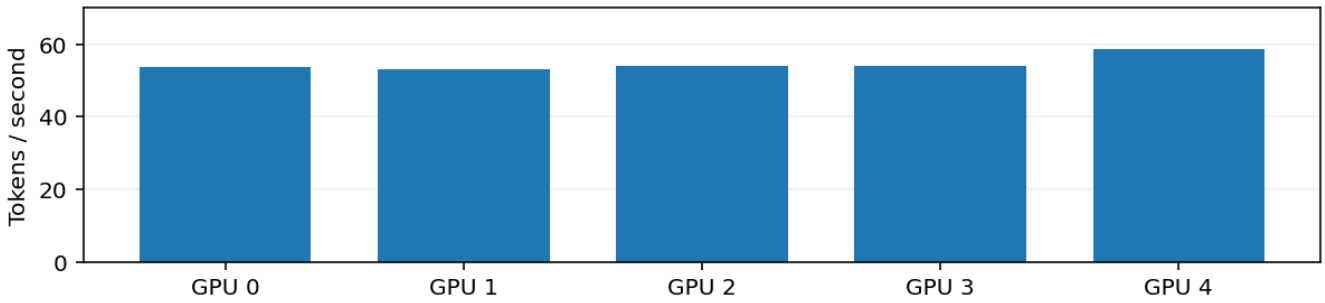
This study examines whether PCIe x1 connectivity is practical for concurrent, GPU-resident local AI inference, compared with an x16-connected GPU. Five independent llama.cpp servers were pinned to NVIDIA GPUs by CUDA UUID. Each model received three concurrent LONG requests of 512 generated tokens per GPU. Three rounds of simultaneous server startup-to-ready measurements were also recorded for each model.

GPU	Graphics processor	PCIe link
0	GTX 1660 Ti · 6 GB	x1
1	GTX 1660 SUPER · 6 GB	x1
2	GTX 1660 SUPER · 6 GB	x1
3	GTX 1660 SUPER · 6 GB	x1
4	GTX 1660 Ti · 6 GB	x16

Models: Gemma 4 E4B Heretic Q4_K_M and Qwen3-8B Q4_K_M. The comparison is observational, not a controlled PCIe-width-only experiment.

2. Gemma 4 — LONG 512-token inference

GPU	Run 1	Run 2	Run 3	Mean tok/s
0	53.98	53.52	53.21	53.57
1	54.13	53.43	52.11	53.22
2	54.25	53.62	53.83	53.90
3	54.50	53.73	53.80	54.01
4	58.48	58.81	58.80	58.70



GPU 4 (x16) reached 58.70 tok/s; GPUs 0–3 (x1) reached 53.22–54.01 tok/s.



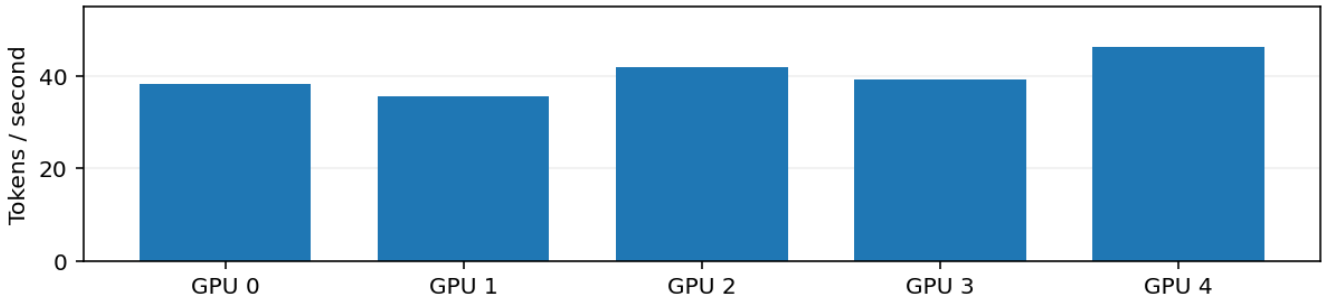
PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

3. Qwen3-8B — LONG 512-token inference

GPU	Run 1	Run 2	Run 3	Mean tok/s
0	42.28	38.46	34.10	38.28
1	39.92	35.62	31.74	35.76
2	43.41	42.17	40.44	42.01
3	41.89	39.24	36.85	39.33
4	46.14	46.52	45.96	46.21



GPU 4 (x16) reached 46.21 tok/s; x1 GPUs reached 35.76–42.01 tok/s. Successive Qwen requests declined on GPUs 0–3.

4. Indicative aggregate generation rates

Model	Sum of five GPU mean rates
Gemma 4 E4B Heretic	273.40 tok/s
Qwen3-8B	201.59 tok/s

These totals are sums of per-GPU means, not independently measured sustained end-to-end throughput.

5. Model-readiness measurement method

Each round launched five model servers concurrently and measured elapsed time from process launch until HTTP /health returned ready. Servers were stopped and GPU memory release verified before the next load. Three Gemma 4 → Qwen3-8B sequences yielded 30 individual readiness measurements. Shutdown and memory-release time are excluded. Round 1 is a cold server start, not necessarily a cold filesystem-cache read.



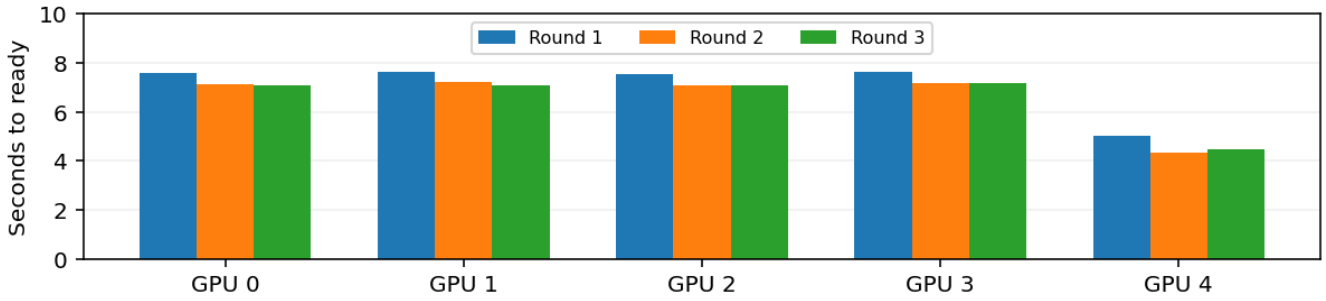
PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

6. Gemma 4 model readiness — three rounds

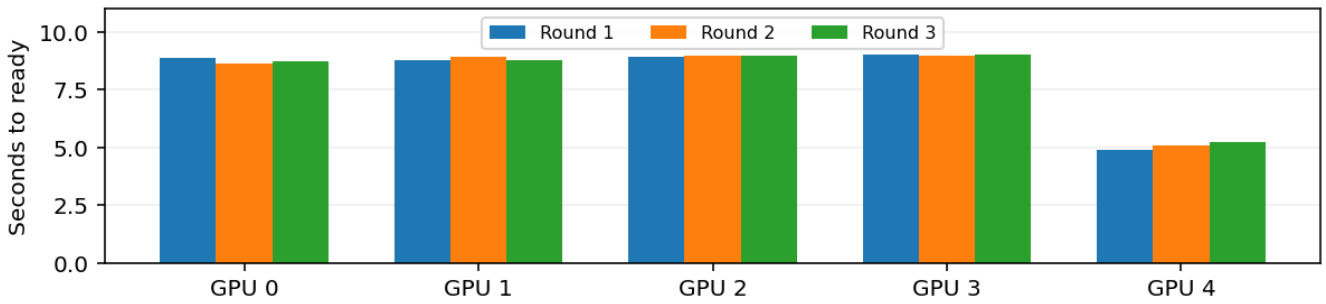
GPU	Round 1 (s)	Round 2 (s)	Round 3 (s)	Mean (s)
0	7.575	7.123	7.088	7.262
1	7.638	7.217	7.088	7.314
2	7.559	7.095	7.095	7.250
3	7.610	7.188	7.189	7.329
4	5.040	4.356	4.460	4.619



GPU 4 averaged 4.619 seconds; x1 GPUs averaged 7.250–7.329 seconds.

7. Qwen3-8B model readiness — three rounds

GPU	Round 1 (s)	Round 2 (s)	Round 3 (s)	Mean (s)
0	8.855	8.625	8.751	8.744
1	8.762	8.917	8.761	8.813
2	8.925	8.987	8.962	8.958
3	9.022	8.952	9.002	8.992
4	4.907	5.091	5.253	5.084



GPU 4 averaged 5.084 seconds; x1 GPUs averaged 8.744–8.992 seconds.

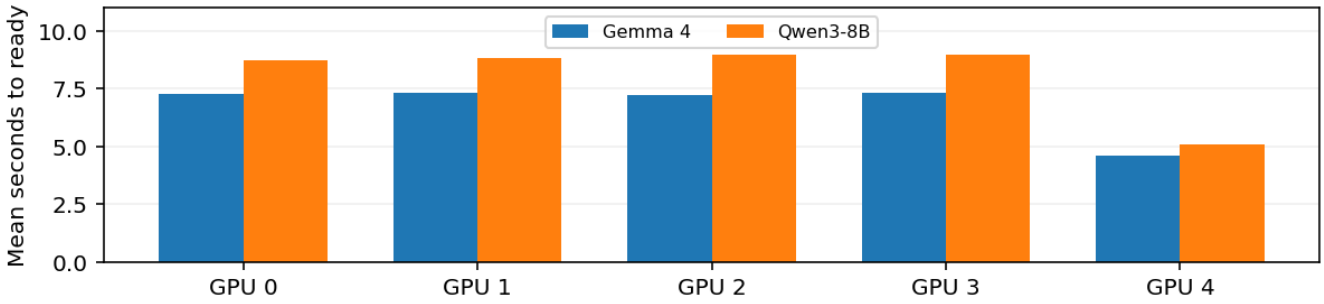


PROFILE

LINKEDIN

AGENTIC AI SYSTEMS

8. Side-by-side readiness comparison



GPU	Gemma 4 mean	Qwen3 mean	PCIe
0	7.262 s	8.744 s	x1
1	7.314 s	8.813 s	x1
2	7.250 s	8.958 s	x1
3	7.329 s	8.992 s	x1
4	4.619 s	5.084 s	x16

9. Technical findings

Concurrent model loading: All 30 server readiness measurements completed successfully. Each model launched concurrently across five GPUs during each round.

GPU 4: The x16-connected GTX 1660 Ti was fastest to ready for both models (4.619 s Gemma; 5.084 s Qwen). This correlation does not independently establish PCIe bandwidth as the cause.

Consistency: Readiness times were relatively repeatable across rounds. The first Gemma load was slower than later loads on every GPU.

Clean-up: Final GPU memory usage returned to initial idle readings: GPU 0, 17 MiB; GPUs 1–4, 7 MiB.

Measurement boundary: Readiness means startup-to-ready, not complete model switching, first-token latency or isolated PCIe transfer time.

10. Conclusion — does PCIe x1 really matter?

PCIe x1 was demonstrably usable for five simultaneous local inference workloads. The x16-connected GPU led in both token generation and model readiness, especially during startup. However, these results alone cannot isolate PCIe width from GPU differences, power/clock behaviour, system topology, caching or other variables. For a causal x1-versus-x16 assessment, retest the same physical GPU under matched settings and workloads at both link widths.

Sources: AIGENESIS terminal benchmark five-gpu-long-clean-20261024-151402; model-readiness benchmark model-switch-20261024-153914. Values rounded for presentation.