



PROFILE



LinkedIn
LINKEDIN



AGENTIC AI SYSTEMS

AGENTIC AI SYSTEMS / TECHNICAL PUBLICATIONS

AIGENESIS: Five-GPU Multi-Model and Multi-Inference Architecture

Consolidated hardware topology, PCIe analysis, shared-model 10- and 30-request benchmarks, GPU operating modes, orchestration, decision control and historical performance.

Publication metadata	Value
Author	Reddy (Eddie) Jamisola
Organisation	Agentic AI Systems
Document ID	202610-AASPUB-0007
Publication date	22 October 2026
Scope	AIGENESIS / DarkFactory; FOCUS excluded
Evidence window	Historical measurements and service records, August 2026

Abstract

AIGENESIS is a consumer-hardware multi-GPU inference host combining five NVIDIA GTX 1660-class GPUs with shared-model request concurrency, model distribution, heterogeneous role-specific endpoints and multi-stage AI orchestration. The retained historical measurements include 10- and 30-request shared-model experiments, concurrent five-endpoint review, Foreman and Arbiter control, deterministic memory guardrails, and timestamped BlackBox execution records. A selected six-scenario synthetic workflow recorded 114 matched model requests and responses, 35,054 completion tokens and three accepted scenarios out of six. This paper consolidates the supplied technical records into a single non-duplicative engineering publication.



PROFILE



LinkedIn
LINKEDIN



AGENTIC AI SYSTEMS

Key measured observations

Dimension	Recorded observation
Shared-model concurrency	10 requests: Gemma 15.74 s / 81.33 tok/s; Qwen3-30B 33.65 s / 38.04 tok/s
30-request inference	Qwen3-30B: 72.78 s / 52.76 tok/s; 1.13 s completion spread
Concurrent reviewer workflow	Five overlapping model calls to endpoints 8190-8194
Selected control workflow	114/114 HTTP 200 responses; 35,054 completion tokens; 3 PASS / 3 FAIL
Hardware	Five 6 GB GPUs; four PCIe 3.0 x1 risers, one x16 connection reported



PROFILE



AGENTIC AI SYSTEMS

1. Physical platform and GPU inventory

Component	Recorded specification
Motherboard	Gigabyte X570 UD, AMD X570 chipset (earlier reported inventory; not repeated in 22 Oct preflight)
BIOS	F41b, 14 April 2026 (earlier reported inspection)
CPU	AMD Ryzen 5 2600, 6 cores / 12 threads
System RAM	37 GiB observed 22 Oct; 76 GiB reported 17 Oct; conflicting snapshots require date-specific attribution
Swap	8 GiB; 8 GiB used on 22 Oct
Storage	Approximately 1.5 TB root logical volume reported previously; physical drives not identified
GPU 0	NVIDIA GeForce GTX 1660 Ti — 6 GB VRAM
GPU 1	NVIDIA GeForce GTX 1660 SUPER — 6 GB VRAM
GPU 2	NVIDIA GeForce GTX 1660 SUPER — 6 GB VRAM
GPU 3	NVIDIA GeForce GTX 1660 SUPER — 6 GB VRAM
GPU 4	NVIDIA GeForce GTX 1660 Ti — 6 GB VRAM
VRAM total	30 GB nominal across five devices; not unified memory
Interconnect	GPU 0–3 negotiated Gen1 x1; GPU 4 negotiated Gen1 x16 (22 Oct); maximum capability Gen3 x16 reported for all GPUs. Riser configuration reported as four x1, one x16
PCI bus addresses	GPU 0: 04:00.0; 1: 05:00.0; 2: 07:00.0; 3: 08:00.0; 4: 0C:00.0 (earlier inventory)
Topology	GPU 0–3 PIX, GPU 4 PHB (earlier NVIDIA topology output)
Platform	Ubuntu, systemd services, llama.cpp endpoints, DarkFactory API / control services

2. PCIe topology and transfer limits

The 22 October NVIDIA snapshot showed Gen1 as the CURRENT negotiated generation on every GPU, not Gen3. GPUs 0–3 were current x1; GPU 4 current x16. Gen1 theoretical payload is approximately 0.250 GB/s for x1 and 4.000 GB/s for x16 per direction. Gen3 theoretical payload is approximately 0.985 GB/s for x1 and 15.754 GB/s for x16. All are theoretical ceilings, not application transfer measurements. GPU link speeds may downshift while idle, so the snapshot does not establish the generation during an active DMA transfer.



PROFILE

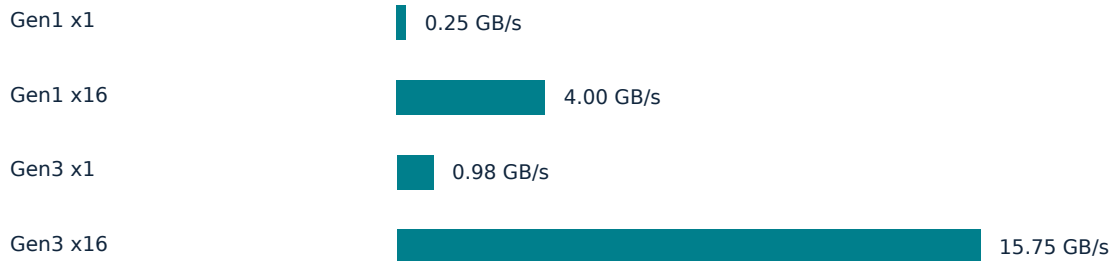


LINKEDIN



AGENTIC AI SYSTEMS

Figure 1. Theoretical PCIe payload ceilings by link generation and width



A 4 GB transfer requires at least approximately 16.0 seconds at ideal Gen1 x1 capacity, 1.0 second at Gen1 x16, 4.06 seconds at Gen3 x1, or 0.254 seconds at Gen3 x16, excluding disk reads, DMA overhead, allocation, synchronization and model initialization. A GPU-resident inference may be less link-sensitive than repeated model loading or CPU/RAM offloading. Four independent x1 links do not become one x4 link for a single device.

3. GPU deployment and inference operating modes

Mode	Arrangement	Evidence status
A — single inference	One model, one request, one GPU	Architecture configuration
B — shared-model inference	One loaded model serves overlapping requests / batching slots	10- and 30-request historical benchmarks
C — one model / five GPUs	One large model distributed across GPU 0-4	Qwen3-30B and Qwen3.8-27B configurations recorded
D — multi-model split	Distinct models on different GPU groups	Historical heterogeneous service profiles
E — five reviewers	Five Qwen3-8B reviewer/digest endpoints	Peak five overlapping requests in BlackBox
F — 3+2 GPU partition	One model on three GPUs, another on two	Recorded allocation design; isolated timing unavailable
G — control-heavy	Foreman GPUs 0-1; Arbiter GPUs 2-4	Historical profile inventory
H — CPU/RAM offload	Host-resident layers or KV data plus GPU compute	No isolated offload throughput result
I — dynamic mode switch	Stop previous profiles, launch target models, await health	Sequence recorded; total reload time unavailable



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

4. Model and service inventory

This inventory is reconstructed from the 22 October model configuration files, running processes and historical benchmark profiles. A model file, configured service and running service are distinct states. Ports identify endpoints, not physical GPUs. The 22 October snapshot had 8103 and 8104 in auto-restart following GPU allocation failures.

Model / quantization	Endpoint	Role / recorded usage
Qwen3-30B-A3B Q4_K_M	8104 / 5-GPU profiles	Large MoE writer/recovery endpoint; 10/30 request experiments
Gemma-4-26B-A4B	8108	Writer; historical shared-model 4-5 and 10-request tests
Qwen3-8B Q4_K_M (five instances)	8190-8194	Five reviewer / digest agents; configured -ngl 999, -c 16000, -np 1, q4_0 KV cache; physical GPU affinity not verified
Llama 3.1 8B Q4_K_M	8201	Foreman; low-latency control decisions
DeepSeek-R1-Distill-Qwen-14B Q4_K_M	8202	Arbiter; ACCEPT / REDO / HUMAN_REVIEW decisions
Qwen3-4B	8210	Gatekeeper
Dolphin-Mistral-24B Venice Q4_K_M	8103	Configured five-GPU model; 22 Oct OOM restart loop; role not confirmed
Qwen3.8-27B Q6_K_XL	8110 experimental	Large-model five-GPU deployment (~25 GiB weights)
Qwen3-8B Q4_K_M	8101	CPU inference, -ngl 0, 32,768 context
Gemma-4-E4B-Heretic Q4_K_M	8102	CPU inference, -ngl 0, 32,768 context
DeepSeek-R1-Distill-Qwen-14B Q4_K_M	8105	CPU inference, -ngl 0, 32,768 context
Phi-4 Q4_K_M	8106	CPU inference, -ngl 0, 32,768 context
Llama 3.1 8B Q4_K_M	8107	CPU inference, -ngl 0, 32,768 context
Qwen3-Coder-30B-A3B Q4_K_M	model file only	17.28 GiB GGUF observed; active serving role not confirmed
Qwen3-VL-30B-A3B Q4_K_M	model file only	17.28 GiB GGUF observed; active serving role not confirmed



PROFILE



Linked in
LINKEDIN



AGENTIC AI SYSTEMS

5. Shared-model concurrent inference benchmarks

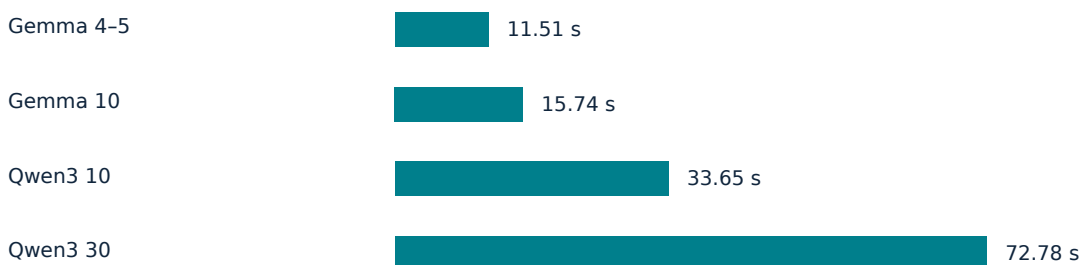
These experiments count simultaneous inference requests served by model processes; they do not represent 10 or 30 distinct GPUs or 30 separately loaded models. Aggregate throughput is total generated output tokens divided by measured elapsed time.

Model / run	Requests	Elapsed	Output tokens	Aggregate rate
Gemma shared-model	4-5	11.51 s	Not retained	55.61 tok/s
Gemma shared-model	10	15.74 s	~1,280	81.33 tok/s
Qwen3-30B slots	10	33.65 s	~1,280	38.04 tok/s
Qwen3-30B agents	30	72.78 s	~3,840	52.76 tok/s

Figure 2. Aggregate generation throughput



Figure 3. Recorded benchmark elapsed time



The Qwen3-30B 30-request experiment recorded a 1.13-second completion spread. A historical four-slot / ten-request Gemma test reported 10/10 successful requests, 1,280 tokens and 15.74 seconds; the four-slot configuration must not be confused with ten model slots. Comparisons across model families require



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

matching prompts, quantization, slot settings and hardware allocation.



6. Context, batching and model switching

One recorded Qwen3-30B-A3B Q4_K_M llama.cpp configuration used -np 10 with 32,768 total context tokens and continuous batching (historically described as approximately 3,328 tokens per slot). Another recorded configuration changed -np 30 to -np 1. An experimental single-slot deployment used a qwen3moe.context_length metadata override to 65,536 and linear RoPE scale 1.6, reporting n_ctx_slot=65,536.

A roughly 22-second service startup was recorded for that experimental configuration. It is not a measured end-to-end mode switch: a full switch may also include stopping old services, releasing VRAM, loading weights, health checks and routing changes.

7. DarkFactory control-plane architecture

The selected story-memory workflow invokes five reviewer endpoints concurrently, then a Foreman, a Qwen3-30B-class endpoint, a second reviewer batch, a second Foreman call and an Arbiter. Rework scenarios add further passes. Deterministic guardrails can confirm or override a raw Arbiter decision.

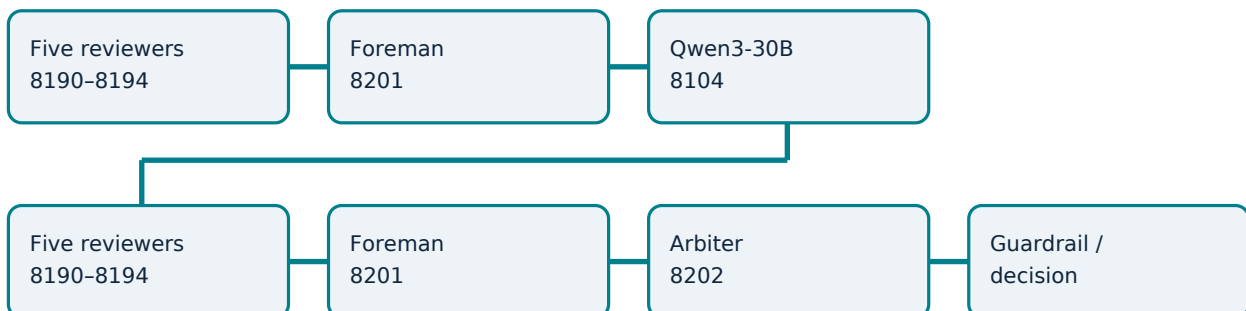


Figure 4. Observed first-pass endpoint ordering; additional redo passes vary by scenario.



8. Detailed first-scenario execution timeline

Phase	Start offset	End offset	Model call wall time
Reviewer batch 1, 5 parallel	+0.00 s	up to +18.66 s	16.73-18.66 s
Foreman 1 / 8201	+39.30 s	+39.98 s	0.68 s
Qwen3-30B / 8104	+74.68 s	+136.37 s	61.69 s
Reviewer batch 2, 5 parallel	+146.52 s	up to +180.77 s	29.92-34.24 s
Foreman 2 / 8201	+199.72 s	+200.62 s	0.90 s
Arbiter / 8202	+200.62 s	+294.62 s	94.00 s

Observed inter-stage gaps: first reviewer completion to Foreman 1 = 20.64 s; Foreman 1 to Qwen start = 34.70 s; Qwen completion to second reviewer start = 10.15 s; second reviewer completion to Foreman 2 = 18.95 s. These are handoff intervals, not measured model-unload or reload times.



PROFILE



LinkedIn
LINKEDIN



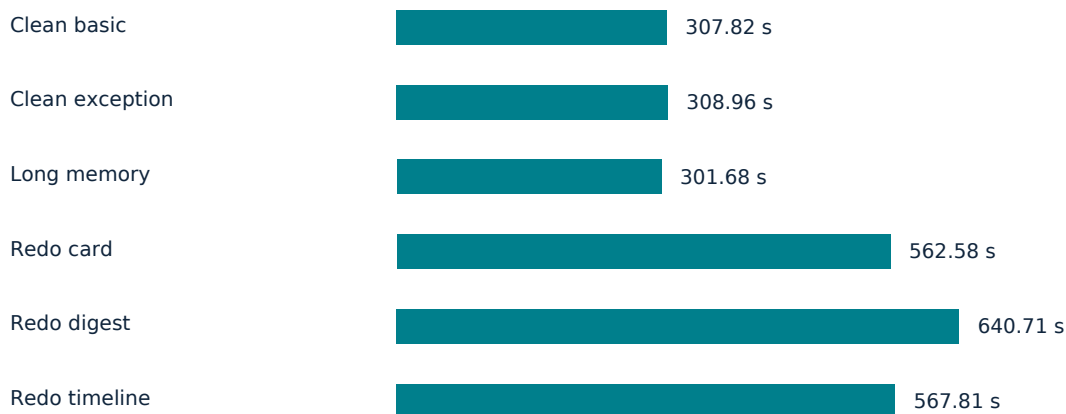
AGENTIC AI SYSTEMS

9. Six-scenario synthetic control benchmark

The selected historical run is story-memory-gpu-control-20260812-205057. Model requests were paired with responses using call identifiers. Scenario intervals end at the next scenario start; the final scenario ends at its last model response. All six reached peak five-way request concurrency.

Scenario	Elapsed	Calls	Actual decision	Result
CLEAN_BASIC	307.819 s	14	REDO_OUTPUT	FAIL
CLEAN_EXCEPTION	308.956 s	14	REDO_OUTPUT	FAIL
CLEAN_LONG_MEMORY	301.678 s	14	ACCEPT	PASS
REDO_OUTPUT_CARD	562.584 s	22	REDO_OUTPUT	FAIL
REDO_DIGEST_MEMORY	640.706 s	28	REDO_DIGEST	PASS
REDO_OUTPUT_TIMELINE	567.808 s	22	REDO_OUTPUT	PASS

Figure 5. Scenario elapsed intervals



The three clean scenarios averaged 306.151 s; the three redo scenarios averaged 590.366 s, or 1.93 times the clean-group interval. Because scenario content differs, the ratio is descriptive and does not isolate retry overhead.



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

10. Agent-call breakdown and overlapping compute

Scenario group	8104	Each 8190-8194	8201	8202	Total
Each clean scenario	1	2	2	1	14
Each REDO_OUTPUT scenario	2	3	3	2	22
REDO_DIGEST_MEMORY	2	4	4	2	28

Scenario	Sum of model-call wall time	Workflow interval
CLEAN_BASIC	407.455 s	307.819 s
CLEAN_EXCEPTION	432.316 s	308.956 s
CLEAN_LONG_MEMORY	418.790 s	301.678 s
REDO_OUTPUT_CARD	774.679 s	562.584 s
REDO_DIGEST_MEMORY	884.804 s	640.706 s
REDO_OUTPUT_TIMELINE	768.808 s	567.808 s

Summed model-call wall time exceeds workflow elapsed time because model requests overlap. The selected run recorded 114 requests, 114 HTTP 200 responses and 35,054 completion tokens; the 50% scenario acceptance rate is a different reliability metric from HTTP response success.



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

11. Decision integrity, arbitration and memory guardrails

REDO_OUTPUT_CARD recorded two raw REDO_OUTPUT Arbiter decisions and a STRONG_REVIEW_CONSENSUS guardrail confirmation, yet the stored scenario result was FAIL. The underlying acceptance rule was not recovered from the summary alone. REDO_DIGEST_MEMORY recorded a raw REDO_OUTPUT decision overridden to REDO_DIGEST by TERMINAL_MEMORY_STATE_CONFLICT and passed. REDO_OUTPUT_TIMELINE recorded raw REDO_OUTPUT and ACCEPT decisions, with recorded actual REDO_OUTPUT and PASS. Raw model decision, guardrail outcome and scenario result must therefore be treated as separate fields.

12. Historical BlackBox run inventory

Run ID suffix	Requests / responses	Recorded span	Event / outcome
20260811-225405	70 / 70	22m 33.744s	ValueError: empty Arbiter control token
20260812-003946	0 / 0	35.492s	SystemExit(130); interrupted
20260812-004136	84 / 84	27m 53.879s	2 PASS / 4 FAIL; SystemExit(2)
20260812-053428	114 / 114	43m 45.401s	3 PASS / 3 FAIL; SystemExit(2)
20260812-194643	114 / 114	44m 37.253s	3 PASS / 3 FAIL; SystemExit(2)
20260812-205057	114 / 114	45m 21.438s	3 PASS / 3 FAIL; SystemExit(2)
20260813-012014	26 / 22	8m 25.351s	Incomplete; four requests unmatched

Across seven inspected logs: 522 requests, 518 responses; six logs recorded FATAL_EXCEPTION events. Four completed six-scenario summaries reported 11 PASS and 13 FAIL (45.83%). These recorded spans include telemetry and cleanup and should not be presented as independently confirmed successful production job durations.

13. Separate writing and editorial observations

Workload	Recorded observation
Novel chapter / Qwen3-30B	2,994 completion tokens; 3,758 prompt tokens; 4,956 words; 27.104 generation tok/s; one expansion round
Editorial first writer / Qwen3-30B-A3B Q4_K_M	892 words; 1,956 completion tokens; 460 prompt tokens; 38.534 s; 60.996 generation tok/s
Editorial pipeline code	Four conceptual stages: first writer, second writer, deep editor, final editor; only first-writer stage confirmed in retained test manifest



PROFILE



LinkedIn

LINKEDIN



AGENTIC AI SYSTEMS

14. Source and methodology

This document consolidates the four supplied AIGENESIS PDFs, six retained terminal transcripts and historical terminal-derived BlackBox summaries, preserving the distinct 10/30 shared-model experiments, five-reviewer control test, hardware configuration and writer/editorial observations. The principal log source is DarkFactory data/gpu-tests/story-memory-gpu-control-20260812-205057/blackbox/flight-recorder.log with its summary.json. Related code paths identified historically include app/novel_engine.py, app/story_factory/production_pipeline.py, app/story_factory/cleanup_pipeline.py, app/story_factory/beat_runner.py, editorial/pipeline.py, scheduler/controller.py and scripts/darkfactory_supervisor.py.

All reported timings are historical observations or explicitly labelled calculations. The record does not establish an isolated x1-versus-x16 application benchmark, exact model switch reload duration, or a direct controlled single-agent versus five-agent speedup. No new inference workload was executed in preparing this publication.



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

15. Hardware state, GPU utilization and model residency

The 22 October 2026 07:25 ACST read-only preflight reported 37 GiB RAM total, 34 GiB used, 951 MiB free, 3.3 GiB available, and 8.0 GiB of 8.0 GiB swap in use. These values reflect a highly pressured snapshot; they do not establish that the same memory configuration was installed during August benchmarks. A prior inspection was reported as 76 GiB on 17 October. The discrepancy is retained, not silently reconciled.

GPU	VRAM capacity	VRAM used at snapshot	Utilization	Negotiated link
0 — GTX 1660 Ti	6,144 MiB	4,978 MiB	0%	Gen1 x1
1 — GTX 1660 SUPER	6,144 MiB	3,763 MiB	0%	Gen1 x1
2 — GTX 1660 SUPER	6,144 MiB	3,763 MiB	0%	Gen1 x1
3 — GTX 1660 SUPER	6,144 MiB	4,039 MiB	0%	Gen1 x1
4 — GTX 1660 Ti	6,144 MiB	3,789 MiB	0%	Gen1 x16

Figure 6. GPU memory occupancy in 22 October snapshot



The earlier PCI bus IDs and topology were: GPU 0 at 04:00.0, GPU 1 at 05:00.0, GPU 2 at 07:00.0, GPU 3 at 08:00.0, GPU 4 at 0C:00.0; GPU 0–3 PIX and GPU 4 PHB. The observed Gen1 state is a link-training snapshot, not a sustained workload transfer-rate benchmark. GPU 4 was the only card observed with a current x16 width.



PROFILE



Linkedin
LINKEDIN



AGENTIC AI SYSTEMS

16. Complete model file inventory and deployment states

GGUF model	Observed file size	Deployment evidence
Qwen3.8-27B UD Q6_K_XL	24.14 GiB	Experimental five-GPU / port 8110 reported
Qwen3-Coder-30B-A3B Instruct Q4_K_M	17.28 GiB	Local model file; serving unverified
Qwen3-VL-30B-A3B Instruct Q4_K_M	17.28 GiB	Local model file; serving unverified
Qwen3-30B-A3B Q4_K_M	17.28 GiB	8104; multi-GPU profile; October startup failures
Gemma-4-26B-A4B Q4_K_M	15.64 GiB	8108; primary writer, active in snapshot
Dolphin-Mistral-24B Venice Q4_K_M	13.35 GiB	8103 configured; startup OOM
Phi-4 Q4_K_M	8.43 GiB	8106 CPU service
DeepSeek-R1-Distill-Qwen-14B Q4_K_M	8.37 GiB	8105 CPU; 8202 historical Arbiter
Gemma-4-E4B-Heretic Q4_K_M	4.94 GiB	8102 CPU service
Qwen3-8B Q4_K_M	4.68 GiB	8101 CPU; 8190-8194 reviewer profiles
Llama 3.1 8B Q4_K_M	4.58 GiB	8107 CPU; 8201 historical Foreman
Qwen3-4B Q4_K_M	2.33 GiB	8210 CPU Gatekeeper

Additional ggml-vocab-*.gguf files were present; these are vocabulary assets, not independent full model checkpoints, and are excluded from the model count. Model weights on disk cannot be summed as concurrently GPU-resident allocations.



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

17. Inference server settings and physical GPU allocations

Profile / port	Configured runtime parameters
8108 writer	-ngl 999; -c 32768; -b 256; -ub 64; -fa on; --split-mode layer; --tensor-split 1,1,1,1,1; q8_0 K/V cache; -t 16
8104 Qwen30B	-ngl 999; -c 65536; -np 1; layer split 1,1,1,1,1; RoPE linear scale 1.6; context-length override 65536
8103 Dolphin	-ngl 999; -c 32768; layer split 1,1,1,1,1; q8_0 K/V cache
8190-8194 reviewers	Each: -ngl 999; -c 16000; --parallel 1; q4_0 K/V cache; -b 256; -ub 64; flash attention
8201 Foreman	-ngl 999; -c 32768; --parallel 1; layer split 1,1; q4_0 K/V cache
8202 Arbiter	-ngl 999; -c 32768; --parallel 1; layer split 1,1,1; q4_0 K/V cache
8101, 8102, 8105-8107	-ngl 0 CPU profiles; -c 32768; 16 CPU threads (historical configured settings)
8210 Gatekeeper	--n-gpu-layers 0; --ctx-size 8192; 8 CPU threads; q8_0 K/V cache

The October Gemma writer was a five-GPU layer-split process; this does not imply that the historical reviewer profile and Gemma writer could simultaneously fit alongside other five-GPU models. Separate profile transitions are essential for interpreting model counts and capacity.



PROFILE



LinkedIn
LINKEDIN



AGENTIC AI SYSTEMS

18. Switching profiles, startup and failure timing

Dynamic switching was implemented as a control sequence: request target profile; stop existing GPU services; wait for GPU pool clearance; start target model; wait for model health; resume work or recover. Supervisor logs on 15–16 August record both successful and failed transitions. A successful 15 August event switched toward the Qwen3-30B recovery analyst at 00:54:53.628 and logged READY at 00:55:22.614: approximately 28.985 seconds between those two events. This is an observed event-to-event transition interval for that instance, not a general cold-load guarantee.

Event / test	Observed interval or result	Interpretation
15 Aug recovery switch → READY	~28.985 s	One successful supervisor event-to-ready interval
16 Aug recovery attempts	~241 s from switch start to pool-clear failure in repeated samples	GPU pool did not become clear; NOT a successful mode switch
Experimental Qwen30B single-slot startup	~22 s reported	Server startup; not full profile-switch latency
22 Oct Dolphin 8103 service failure	~23.5 s systemd wall per failed start	GPU0 CUDA allocation OOM, restart loop
22 Oct Qwen30B 8104 failure	~12.5 s internal model-load attempt before allocation failure	Failed CUDA allocation; not a successful startup
First scenario reviewer→Foreman gap	20.64 s	Orchestration handoff, not model loading
First scenario Foreman→Qwen gap	34.70 s	Orchestration handoff, not model loading

The 22 October preflight found the Gemma writer running while Dolphin and Qwen30B services were restarting. Dolphin failed allocating approximately 2,938.71 MiB on CUDA device 0; Qwen30B failed allocating approximately 3,699.72 MiB on CUDA device 0. This is direct evidence of GPU memory contention and failed co-residency, not a successful dynamic-switch benchmark.

19. Shared-model concurrency scaling and slot distinctions

The 4–5, 10 and 30 concurrent-request figures were collected in separate historical model-serving experiments. One historical 10-request result used only four serving slots, completing 10/10 requests and generating 1,280 tokens in 15.74 seconds. Thus the request count and server slot count must be reported separately. Another Qwen3-30B configuration used -np 10 with total context 32,768, and a 30-request mode used -np 30. A later transition from -np 30 to -np 1 was recorded; this is a change of concurrency profile, not an additional model.



Test	Requests	Elapsed	Aggregate throughput	Completion detail
Gemma 4-5	4-5	11.51 s	55.61 tok/s	Aggregate tokens not retained
Gemma 10	10	15.74 s	81.33 tok/s	~1,280 tokens; 10/10 completed
Qwen3-30B 10	10	33.65 s	38.04 tok/s	~1,280 tokens
Qwen3-30B 30	30	72.78 s	52.76 tok/s	~3,840 tokens; 1.13 s finish spread

Figure 7. Concurrent-request throughput by recorded model profile



The 30-request run has 3x as many requests as the 10-request run and approximately 1.57x the aggregate throughput within the Qwen profile (52.76 / 38.04); latency also rises from 33.65 s to 72.78 s. This is a descriptive cross-run ratio, not a controlled scaling law. Per-request latency distributions and exact prompts were not retained in the supplied consolidated records.

20. Reliability and supervisor behavior

The control pipeline logged decision extraction, strong-consensus confirmation, memory-conflict override, scenario context changes and telemetry heartbeats. The seven historical BlackBox traces contained 522 requests and 518 responses; the four unmatched requests were in an incomplete run. Four complete six-scenario summaries totaled 11 PASS / 13 FAIL. Supervisor logs also document repeated GPU-pool-clear failures during recovery attempts. HTTP 200 responses and job-level success are different metrics.



PROFILE



LINKEDIN



AGENTIC AI SYSTEMS

21. Evidence ledger and reproducibility index

Evidence sources are historical terminal transcripts and the four predecessor PDFs. The publication separates source-reported facts from calculations and unresolved fields. The table below names the exact evidence artifacts so an independent reviewer can reproduce the inventory without inventing new measurements.

Evidence artifact	What it substantiates
Pasted text (20261021-215535).txt	22 Oct CPU/RAM/GPU/PCIe snapshot, processes, model configs, GGUF file sizes
Pasted text (20261021-215629).txt	8103/8104 startup failure logs, CUDA allocation failures and service restarts
Pasted text (20261021-215938).txt	Supervisor mode-switch history, 15 Aug READY event, 16 Aug GPU-clear failures
Pasted text (20261021-220428).txt	DarkFactory file and workflow discovery
Pasted text (20261021-220516).txt	Orchestrator, editorial and agent-role code references
Pasted text (20261021-222856).txt	Six-scenario historical summaries and control-heavy model timing
BlackBox run story-memory-gpu-control-20260812-205057	114 request/response pairs, scenario timing, decision extraction, guardrail events
Earlier four AIGENESIS PDFs	Previously compiled inference tests, figures, hardware and role descriptions
Prior hardware inventory reported in conversation	Gigabyte X570 UD, BIOS F41b, PCI addresses, 76 GiB RAM and root logical volume; cross-check pending

22. Explicitly unresolved inventory fields

The available retained evidence does not establish PSU manufacturer/wattage, exact DIMM part numbers or speed, physical storage drive manufacturer/model, live PCIe throughput under sustained load, per-model power draw, exact model-to-GPU affinity for all historical profiles, or the complete model-switch timing distribution. The motherboard and BIOS identifiers derive from an earlier reported inspection rather than the attached 22 October preflight. No values have been invented to fill these gaps.



PROFILE



LinkedIn
LINKEDIN



AGENTIC AI SYSTEMS

23. Engineering interpretation

The system is best understood as a set of selectable deployment profiles, not as every model loaded on all five GPUs at once. Distinct validated behaviors include: five concurrent reviewer HTTP requests; 10/30 request shared-model serving; five-GPU layer-split writer configuration; separate CPU resident model services; control-model arbitration and deterministic guardrails; and supervisor-driven recovery mode changes. The October GPU pressure and repeated allocation failures illustrate the practical limit of simultaneous residency on five 6 GB cards.

15. Conclusion

AIGENESIS combines heterogeneous PCIe-connected consumer GPUs with several distinct inference arrangements: large-model multi-GPU distribution, shared-model multi-request serving, five-way reviewer execution and role-specific control models. Recorded 10- and 30-request tests establish aggregate shared-model throughput, while BlackBox traces establish overlapping reviewer calls, staged Foreman and Arbiter decisions, deterministic memory guardrails and measurable rework overhead. The combined evidence characterizes a concrete implemented engineering system and its observed performance, including recorded scenario failures and exception events.