

Best use of Nebius: agreed implementation plan

Three agents with different stakes each proposed an implementation, then voted on a merged draft:

- **Platform architect:** depth and correctness of Nebius usage.
- **Judge advocate:** score with these specific judges.
- **Pragmatic builder:** ships in 23 days, and every claim is real.

This document is the result, with every accepted amendment folded in. Where it differs from [ARCHITECTURE.md](#), this plan wins.

Day 1 = Wed Oct 7, 2026. Submission deadline: Fri Oct 30, 10:00 PDT.

The one thing we're building

One fully real closed loop. On Day 1, Ultra decides an incident. That night, Night Shift runs for real on Nebius. On Day 2, the promoted Nano handles a held-out variant it never saw. Everything else serves this loop.

STAGE	NEBIUS SERVICE	WHAT RUNS
Day loop	Token Factory (Ultra) + Serverless Endpoint (Nano)	Nano decides; low confidence or safety-critical goes to Ultra. Peers swap precedents with no model call.
Logging	Token Factory Data Lab + Object Storage	Data Lab captures Ultra calls; our own decisions JSONL goes to the bucket, so we don't depend on Data Lab.
1 Replay	Token Factory batch (/v1/bat ches)	Ultra re-judges every Nano decision; disagreement rate sets tomorrow's τ .
2 Dream	Serverless Jobs (CPU fan-out)	Ultra writes K variants per salient episode; shards run the headless sim + planner.
3 Thymus	Serverless Jobs	One gate job, two suites: variants must improve, normal operations must not regress. Shown on screen as the thymus funnel.
4 Consolidate	Serverless Jobs (GPU, H200) + Managed MLflow	LoRA on Nano; MLflow lineage; promote only if it beats today's Nano.
Serving	Serverless Endpoint (vLLM)	One endpoint serves production and candidate Nano on the same base.

STAGE	NEBIUS SERVICE	WHAT RUNS
Perception	Token Factory dedicated endpoint (<code>dedicated/nvidia/Cosmos3-Super-Reasoner-g_EjNm</code> , see config/models.yaml)	Cosmos describes a hazard on first sighting; peers spread it with no further calls.
Plumbing	MysteryBox, Container Registry, one bucket	Secrets, one image for every job, all artifacts.

Decisions

1. **Core story:** the closed loop above.

2. Models:

- Read the IDs from `GET /v1/models?verbose=true` on day 1 and write them into one `models.yaml` . Every call and MLflow run reads from it.
- Choose Ultra on both availability and a day-1 latency/throughput test. Prefer Nemotron 3 Ultra; otherwise use Llama-3.1-Nemotron-Ultra-253B.
- If Ultra's p95 latency is over 30 s, live escalations go to Nemotron 3 Super and Ultra stays on Night Shift.
- Show the exact model IDs on screen and in the README.

3. Nano choice, decided by a smoke test:

- The test passes only if vLLM on a Serverless Endpoint serves the trained adapter (`--enable-lora`) or the merged weights.
- Train on `gpu-h200-sxm` , and confirm GPU quota on day 1.
- Default is Nemotron 3 Nano 30B (MoE, about 3.5B active).
- **Day-3 checkpoint:** if the 30B isn't green, switch immediately to dense Nemotron Nano 9B/12B v2.

4. **Headroom test (day 1):** 50 variants, comparing Nano zero-shot, Nano plus retrieved precedents, and Ultra.

- If there's no accuracy headroom, the before/after headline becomes Ultra calls per 100 events, \$ per task, and p50 latency, stated honestly.

5. **Logging:** Data Lab plus our own JSONL in Object Storage.

6. Replay batch:

- Batch only promises results within 24 h, so submit a small test batch in week 1 and measure the real turnaround.
- The first real Night Shift batch runs by **day 8**.
- Demo replay uses concurrent per-token calls on the same JSONL and code path, labeled as such.

7. Dream fan-out:

- Shard count is a config value: start at 4, and scale to at least 20 (target about 50) once one shard runs clean.

- The width has to be visible in the console job list.
8. **Thymus:** folded into the promotion gate (two suites), presented as the funnel.
9. **Serving (resolves the architect's blocking objection):**
- The day loop's "production Nano" and the candidate run on the **same vLLM endpoint with the same base weights**: base weights on Day 1, adapter afterwards. Otherwise the before/after isn't comparable.
 - Token Factory's Nano is used only for the headroom test.
 - Test both the LoRA-adapter path and the merged-weights path on day 2.
 - Request Token Factory custom-weights beta in week 1, as a bonus only.
10. **Plumbing:** MysteryBox, Container Registry and one bucket, all named in the README's Nebius section.
11. **Cut:**
- Countercurrent pairing (Option A) as a headline; at most one sweep row.
 - Cosmos-rendered hero clips.
 - Any dependence on custom weights.
12. **Cosmos is back in the core loop (updated Oct 6 after the round-2 vote):**
- Verified in the Token Factory console: `nvidia/Cosmos3-Super-Reasoner` runs as a **dedicated endpoint** (us-central1, 4x H200, vision, \$0.37 per minute). This overturns the round-2 premise that Cosmos wasn't on Token Factory.
 - Use it for first-sighting perception: Cosmos describes the hazard once, and peer handshakes spread it. That is the Option B story with a real NVIDIA Physical AI model behind it.
 - The builder's objection still applies: the 2D sim has no camera frames. So perception runs on rendered top-down frames of the sim plus a small set of real warehouse-hazard photos, and the README states which is which.
 - **Cost control is mandatory:** about \$22/hour while running. Min replicas 0 (or stop the endpoint) outside test and recording sessions, batch perception requests together, and log spend in the cost table.
13. **Stretch goals, gated:**
- Tavily safety-data-sheet lookup (sim-only, about 1 day) if the core loop works by day 14.
 - Jetson only if a board is on hand by day 10.
14. **Video:**
- **At least 60 continuous seconds of modules running (Day loop → Night Shift → Day 2), no slide cutaways.** This is a Physical AI eligibility rule.
 - Real Nebius console footage: the batch job, the fan-out job list, the GPU job, the MLflow run.
 - A live API-call log panel (model ID, latency, \$), and time-lapse with visible timestamps.
 - Every number comes from stored runs, traceable by MLflow run ID and storage path.
15. **FEEDBACK.md and the README "Built on Nebius" section start on day 1.**

- Each feedback item states the problem, the measured number, the workaround and a suggested fix. See [FEEDBACK.md](#).

16. If training fails, say so honestly:

- Night Shift becomes replay + dream + gate + precedent consolidation + τ recalibration.
- The video says "consolidated into fleet memory", not "trained into Nano", with any attempted training run shown in MLflow.

Schedule and cut lines

DAY (DATE)	MUST BE TRUE
1 (Oct 7)	Model IDs, Ultra latency test, headroom test, GPU quota confirmed, test batch submitted, custom-weights access requested, Cosmos endpoint smoke test then stopped
2 (Oct 8)	LoRA adapter vs. merged-weights serving paths both tried
3 (Oct 9)	Checkpoint: 30B LoRA smoke test green, or switch to 9B/12B now
7 (Oct 13)	Day loop working end to end (sim, planner, Nano, gate, Ultra, peers); one manual LoRA run proven
8 (Oct 14)	First real Night Shift batch submitted
12 (Oct 18)	First full Night Shift run end to end (batch → shards → gate → LoRA → endpoint)
14 (Oct 20)	Night Shift automated; real before/after number; τ sweep frontier; stretch-goal decision
21 (Oct 27)	Feature freeze; final real runs; record
22–23 (Oct 28–30)	Video, README, cost table from real billing, submit

Vote record (round 2)

Decision 12 was amended after the vote, based on the verified Cosmos endpoint; the votes below are as cast.

#	ARCHITECT	JUDGE ADVOCATE	BUILDER
1 Core loop	accept	accept	accept
2 Models	amend: <code>models.yaml</code>	accept + IDs on screen	amend: latency test, Super fallback
3 Nano smoke test	amend: must serve, H200	accept	amend: quota day 1, day-3 checkpoint
4 Headroom	amend: Ultra calls/100, \$	amend: Ultra calls/100, \$/task, p50	accept

#	ARCHITECT	JUDGE ADVOCATE	BUILDER
5 Logging	accept	accept	accept
6 Batch	amend: measure turnaround wk 1	accept	amend: first real batch by day 8
7 Fan-out	accept	amend: at least 20, ideally ~50	amend: start 4, scale up
8 Thymus as gate	accept	accept	accept
9 Serving	amend (blocking): same weights both days	accept	amend: test merged path too
10 Plumbing	accept	accept + README	accept
11 Cuts	accept	accept	accept
12 Cosmos	amend: optional Nano Omni	amend: document why + feedback	accept
13 Stretch	accept	accept	accept
14 Video	accept	amend: 60 s continuous (eligibility)	accept
15 Feedback	accept	amend: format + 5 seeds	accept
16 Fallback	amend: explicit on screen	accept + wording change	amend: day-3 checkpoint, e2e by day 12

Blocking objections: one (architect, #9), resolved by adopting its amendment. The judge advocate and builder raised none.

Cost estimate (architect, to be replaced by real billing)

About \$200–300 for development, and under \$15 per demo night: Ultra per-token and batch, one LoRA on an H100/H200, two small endpoints that scale to zero, and CPU fan-out.