

Demo use case: the aisle 4 spill

One sentence: a liquid spill at a blind corner is seen once, understood once by Cosmos on Nebius, spread to the whole fleet without another model call, made safe by Ultra when a person walks in, and learned overnight on Nebius, so the next morning a different robot handles a harder version of it alone.

Why this scenario

- **It's a real, everyday warehouse problem.** Spills and slip hazards happen in every fulfillment center, and the expensive part is the minutes between "someone sees it" and "everyone knows about it".
- **Every Nebius service does visible work,** and each one appears exactly once with a clear job:

SERVICE	ITS JOB IN THE DEMO
Cosmos (dedicated endpoint)	Sees the spill
Nemotron Ultra (Token Factory)	Decides the safety-critical moment
Batch + Data Lab	Replays the day
Serverless Jobs	Dream fan-out, gate, LoRA training
MLflow	Promotion record
vLLM Serverless Endpoint	Serves the improved Nano

- **It shows the safety boundary in action.** The person in the lane is the one event the fleet is never allowed to settle on its own.
- **It ends in a measured before/after,** not a claim: Ultra calls per 100 events and cost per 1,000 tasks, from stored runs.

Setting

Site	Fulfillment center "DC-04", zone B, aisles 1–6
Fleet	24 autonomous mobile robots (12 visible on the floor map)
Shift	Day 1 peak, Night 1, Day 2 morning
Operator surface	The NemoFleet Ops console (live floor, alarms, actions, services, Night Shift, cost, model call log)

Script (3 minutes; acts 1–3 run as one continuous take of at least 60 seconds)

Act 1 · Day 1, live (0:20–1:05)

1. Routine (10 s).

- Robots move. The "Where decisions are settled" panel shows Nano handling most events in about 300 ms.
- The model call log scrolls with S1 rows.

2. First sighting (15 s).

- R14's camera frame shows clear liquid pooling at the aisle 4 blind corner.
- **Cosmos3 Super Reasoner on the Token Factory dedicated endpoint** describes it: "clear liquid, possible solvent, near a forklift crossing".
- It's one real call, visible in the log with its model ID, latency and cost.
- A hazard tag appears on cell 41,7.

3. Spread by handshake (15 s).

- Robots pass each other and swap the precedent.
- Two counters tell the story: "robots that know" climbs 1 → 12, while "model calls to spread it" stays at 0.
- The payload inspector shows the hop count rising.

4. The twist (15 s).

- A picker walks into aisle 4 to clean the spill.
- A person in the lane is always safety-critical, so the gate escalates to **Nemotron Ultra on Token Factory**, whatever the peers say.
- Ultra returns: halt R03, quarantine A4, reroute R01 and R05, alert supervisor.
- The JSON passes the schema check; the planner reroutes.
- A critical alarm appears with Acknowledge and Release zone buttons.

Act 2 · Night 1 on Nebius, time-lapse with timestamps (1:05–1:55)

Real Nebius console footage next to the Night Shift panel:

1. Replay.

- Data Lab exports the day's Ultra calls.
- A Token Factory batch re-judges every Nano decision and flags the confident mistakes.

2. Dream.

- Ultra writes variants of the aisle 4 episode: spill plus forklift, low light, a different aisle, a second robot arriving late.
- The job list fills with dream shards on **Serverless Jobs**.

3. Gate.

- Each candidate lesson must fix its variants without regressing normal operations.
- Most are rejected, and the funnel shows it.

4. Consolidate.

- A LoRA fine-tune of Nano runs on an H200 Serverless Job.
- The **MLflow** run shows it beating yesterday's Nano on the held-out set, and it's promoted on the **vLLM Serverless Endpoint**.

Act 3 · Day 2, live (1:55–2:25)

1. A different robot, R12, reaches the same blind corner. This time there's a spill *and* a forklift crossing, a combination no robot saw on Day 1.
2. R12 runs the promoted Nano (adapter night-07), recognizes the situation at confidence 0.91, and reroutes via aisle 2.
3. **Ultra calls: 0.** The telemetry panel reads "learned in dream #217".

Act 4 · Proof (2:25–2:50)

- The learning curve shows Ultra calls per 100 events falling night over night.
- The frontier chart from the Serverless Jobs sweep compares Nano only, Ultra only, the hybrid, and the hybrid with Night Shift on success, latency and cost per task.

Close (2:50–3:00)

"The edge lives it once. Nebius lives it hundreds of times." Then the repo link and the model IDs.

What is real, what is simulated

PART	STATUS
Every model call (Cosmos, Ultra, Nano), batch, job, LoRA, MLflow record, endpoint	Real, recorded, traceable by run ID
Warehouse, robots, physics, the picker	Simulated (deterministic sim plus A*/CBS planner)
Camera frames	Rendered sim frames, plus a small set of real spill photos for the Cosmos call; the README says which is which
Night	Real runs, shown as a time-lapse with visible timestamps

Fallbacks

- **The fine-tune shows no accuracy gain** (headroom test, plan decision 4). Act 3 stays, and R12 handles it from the promoted precedent memory instead. The narration says "consolidated into fleet memory", and the proof becomes Ultra calls, cost and latency.

- **Cosmos endpoint unavailable on recording day.** Use the recorded Cosmos response from the smoke test, labeled as recorded.
- **Ultra p95 latency is too high to show live.** Live escalations go to Nemotron Super (plan decision 2), and Ultra stays on Night Shift.

Optional beats (only if the plan's day-14 gate is green)

- **Tavily safety data sheet.** Ultra identifies the substance, Tavily retrieves its safety data sheet, and the exclusion radius in the quarantine comes from it. This adds one call to Act 1, step 4.
- **Jetson Orin Nano.** A real camera on a Jetson performs the first sighting in Act 1, step 2.

Success criteria for the recording

- ☐ At least 60 continuous seconds of modules running (eligibility rule)
- ☐ Each Nebius service appears once, doing its job, with its name on screen
- ☐ Every number on screen comes from a stored run (MLflow run ID or storage path in the description)
- ☐ The person-in-lane escalation happens on camera, and the planner, not the model, moves the robots
- ☐ Day 2 shows zero Ultra calls for the unseen variant, or the honest fallback wording