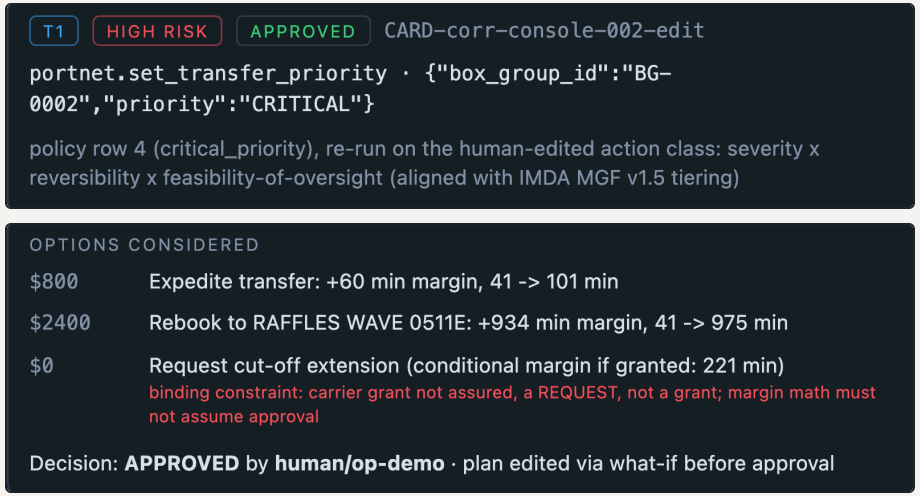
RELAY

RELAY is an exception layer for transhipment connections at a container terminal. It works in the hours between a carrier's schedule change and the next planning cycle.

PSA keeps 84% of its assured vessel calls on time. When one of the rest slips, the first warning is often a carrier's email, and no rules engine acts on free text. RELAY reads that email with a language model, turns it into a checked fact, recomputes which containers will now miss their next ship, has a solver choose the cheapest recovery, and puts each write to a terminal system in front of the duty officer as an approval card that expires into a denial. Every step is recorded on a tamper-evident ledger that can be replayed.



Two parts of one approval card from the RELAY console: the proposed write with the policy row that set its tier, and the options the officer saw, each with its cost and its effect on the margin, followed by the decision. The card was recorded with the pricing gate off, the arm in which this expedite is offered as a card.

AUTONOMY LEVEL CHOSEN

The duty officer stays in the loop. Reads and annotations run unattended; a write to a terminal system waits for a person, and an unanswered card is denied.

DATA

The terminal data is synthetic throughout. Two recorded days of Singapore AIS are used as aggregates. The project holds no PSA data and uses no API keys.

REPRODUCIBLE

Four README commands run offline on a laptop; the full test suite, 1,351 tests, is about eight minutes of that, beside an oracle gate of 28 hand checks and the scripted demo path.

WHERE TO SEE RELAY RUNNING

PROJECT PAGE

`thylinao1.github.io/relay-page`

EXPLANATORY VIDEO, 9 MINUTES

`youtu.be/J3wU0hluJfU`

AUDIO PODCAST, 22 MINUTES

`thylinao1.github.io/relay-page/#listen`

PSA held Assured Port Time on 84% of calls in 2025. For the rest, the warning that matters can arrive as an email.

PSA's 2025 annual report records 518 vessel calls across 17 onboarded services, with 84% meeting their Assured Port Time target. The calls that miss their target include the late inbounds that break transhipment connections: containers that landed from one ship and must be on another before that ship's cut-off. The carrier often knows first, and the carrier's message is prose in an inbox. The structured notice that the terminal operating system acts on can follow hours later.

Rules engines act on structured events. A language model is the one kind of component that can turn the email into a fact, and acting on that fact moves cranes and cargo, so it cannot happen without controls. RELAY limits the model to reading the input; the twin and the solver decide feasibility, and a duty officer approves each write.

FOUR TERMS THE REST OF THE DECK USES

Transhipment connection	A group of containers discharged from one vessel that must load onto another before its cut-off.
Margin	Minutes between the moment the containers can be ready and the cut-off. Below zero the connection is infeasible.
Rollover	A connection that is missed; the containers wait for the next sailing.
The twin	RELAY's deterministic simulator of the terminal, which computes margin and prices actions.

WHAT THE RECORDED AIS SHOWS

Two days of Singapore AIS (Automatic Identification System, the public ship broadcast), 24 and 25 August 2026, pseudonymised, aggregates only. AIS stands in for the structured channel; the recording cannot see terminal events or carrier EDI. When that channel is silent, the carrier's email is the earlier signal.

63 of 74

vessels that berthed 60 or more minutes after their broadcast ETA had given no structured warning of that size beforehand. Most late vessels slipped silently (silent share 0.851, bootstrap 0.770 to 0.919).

463 of 1,106

ETA broadcasts were already in the past when they were sent.

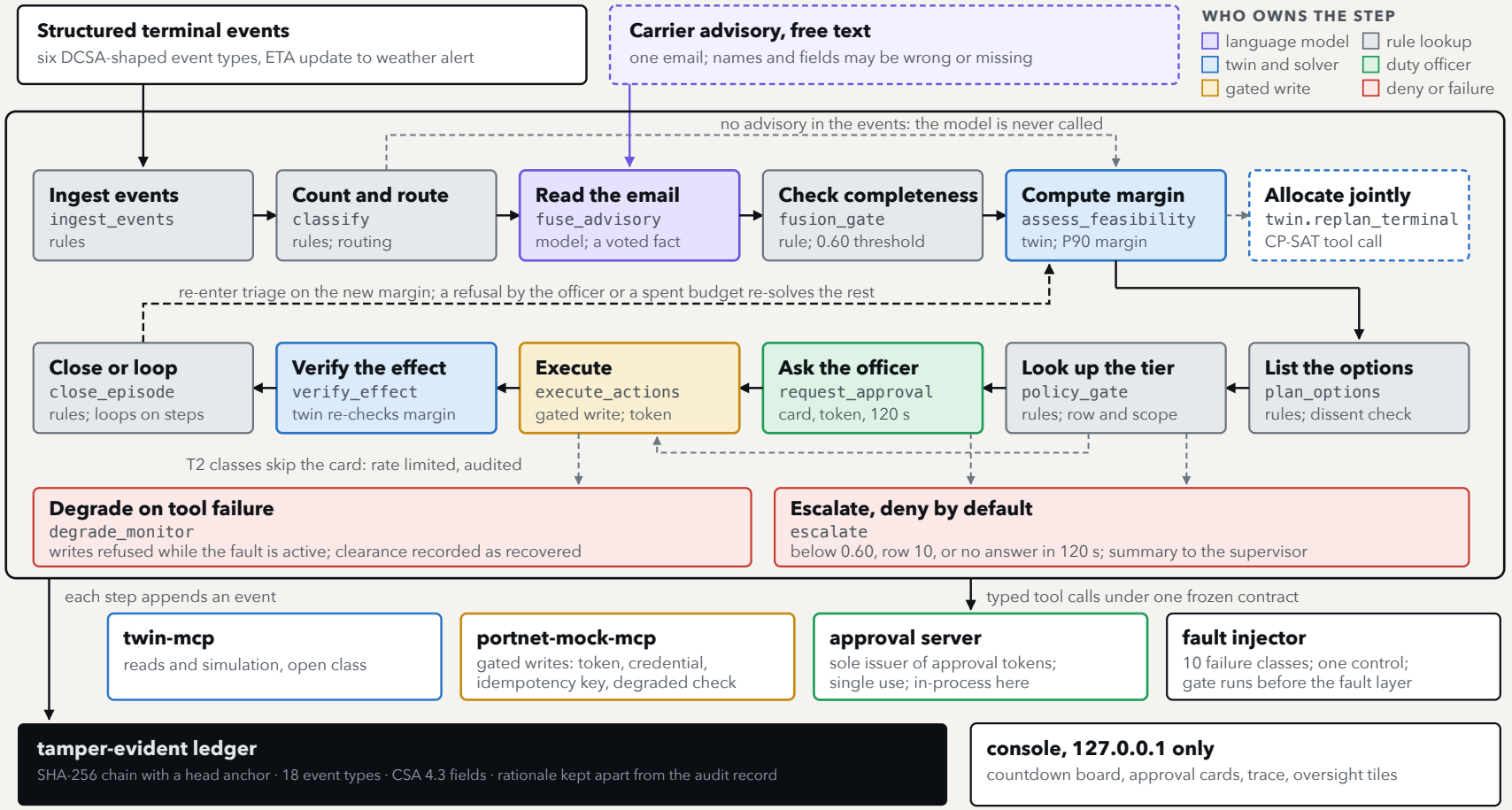
20 of 146

berthed vessels had revised their broadcast ETA by an hour or more before berthing; the other 126 had not. Where a warning did come, the median lead was 282.1 minutes.

PSA Annual Report 2025, Year in Review p.23 · recorded AIS, two Singapore days, pseudonymised

`data/ais_slip.py · evalx/results/ais-slip.json · ais-warning-lead.json`

The model's only job is a checked fact from free text. Deterministic tools decide feasibility, and a policy table decides who is asked.



IMPLEMENTATION

A LangGraph state graph with thirteen nodes, checkpointed in SQLite across the human interrupt, so the graph resumes from the card decision.

Tool modules behind one frozen contract, all in-process in this build: the twin (also exposed as an MCP server over stdio, which the graph does not route through), the terminal-system mock that carries the gated writes (named portnet-mock in the code), the approval server that is the only issuer of approval tokens, and the fault injector.

OR-Tools CP-SAT allocates one action per at-risk connection under the shift budgets. llama3.2:3b runs through Ollama on the local tier; a frontier tier exists and stays off unless an operator sets a key.

Tool choice, tier and action are rule lookups and solver output; the model has no path to any of them. Python 3.11 or later; 1,351 tests; no network needed to reproduce.

Colour marks who owns each step; a dashed edge is a conditional route or a tool call. Measured figures on these slides come from the agent's runs; the console is the operator surface that renders them.

`docs/CONTRACT.md §e, §j · agentcore/graph.py`

Autonomy is set per action class in a table the model cannot edit. An action class the table does not list is denied before a card can exist.

T0
ADVISE ONLY

Berth and berthing-time changes. No write tool exists for them, so the agent can only describe the option and a person takes it elsewhere.

T1
ASK AND APPROVE

Each action receives a card, a single-use approval token bound to its arguments and a 120-second clock. The classes and their per-shift budgets, placeholders the terminal sets in the table: expedite yard transfer (5), CRITICAL priority that preempts other cargo (2), cut-off extension request (3), rebooking proposal (3) and restow order (2). The last four are HIGH risk and need a written justification.

T2
ACT AND AUDIT

Reads and queries of terminal state, risk annotations and internal notifications, twin state ingest, and the escalation summary to the duty supervisor. These run unattended under a rate limit and are audited afterwards.

TWO DEFAULT ROWS

Row 10
AUTO-DENY

An action class not in the table is refused before an approval card exists, and the episode escalates with a written summary.

Row 12
ADVISE_ONLY

An option whose expected value is below its own cost is never proposed as a write. It stays in the record with its three numbers attached and is removed from the solver's candidate set. This is an economic refusal, where T0 is a class rule.

Mirrored in code, with the stricter reading applied where the contract prose is conditional (A2). The tier is a table lookup; the node then runs a deterministic scope check.

WHY THIS LEVEL

The brief says higher autonomy is not automatically better. A write to a terminal system moves cranes and cargo and spends money, the duty officer is on shift, and the cost of an unnecessary write can be measured. A write therefore waits for approval while reads run unattended. When nobody answers within the window, the card is denied and a written summary goes to the duty supervisor. The first deployment proposed on slide 10 is advisory.

TWO WAYS TO DENY BY DEFAULT

A T1 card nobody answers expires after 120 seconds into a denial with a written escalation. An action class with no policy row is denied before a card exists. The demo video shows both paths and the evaluation suite scores both. The 120-second window is the contract's demo value; a per-class window is on the roadmap.

THE OFFICER CAN CHANGE THE PLAN WITHOUT WIDENING THE AGENT'S AUTHORITY

An approver can choose among the solver's enumerated options and can change the priority level. The twin re-simulates the edited plan and the policy gate re-runs on the edited action class, so an expedite raised to CRITICAL moves from row 3 to row 4 and requires a justification. The approval token is then re-bound to the edited arguments.

GOVERNANCE OF THE TABLE

The table lives in the contract and in code; a change to it is a reviewed code change, and the model has no write path to it. The tiers map to the human-involvement levels in IMDA's Model AI Governance Framework for Agentic AI v1.5 and to L0 to L2 in the CSA agentic addendum; L3 is not built. This is an alignment claim, and no certification has been sought.

One email moves a vessel's arrival and leaves CN-0002 with 41 minutes; an edited expedite, approved by the officer, recovers it to 101.

1 · PERCEIVE

Fact from the email

The model parses the free text into a schema-validated fact with allow-listed keys, labelled untrusted. Completeness 0.87 clears the 0.60 gate.

`llm_call · model_rationale · fusion_gate`

2 · INGEST

Twin holds the ETA

The reconciled arrival enters the twin as a vessel fact; both connections on the voyage are updated.

`ingest_fact · vessel_eta_update`

3 · DECIDE

Margin at P90

Margin to cut-off at the P90 of the transfer distribution; CN-0002 is AT_RISK with 41 minutes. The model plays no part in this step.

`feasibility_check AT_RISK 41`

4 · OPTIONS

Three costed options

Expedite the yard transfer, $800, feasible; request a cut-off extension, $0, never feasible alone since a request is not a grant; rebook, $2,400.

`replan_options · 3 options`

5 · GATE

Policy row 3, tier T1

The table lookup sets the tier and raises a card with a single-use token bound to the tool and its arguments.

`policy_gate row 3 · approval_requested`

6 · APPROVE

Officer edits, approves

The officer raises the expedite to CRITICAL; the twin re-simulates, the gate moves to row 4, and the token re-binds.

`human_note approval_card_edited · policy_gate row 4 · approval_granted`

7 · EXECUTE, VERIFY

Margin 41 to 101

The gated write lands, the twin verifies the effect, and close_episode returns to triage on the new margin.

`action_executed · verify_effect · close_episode`

The countdown board before the email and after the approved write: CN-0002 moves from 41 minutes at risk to 101 feasible, and CN-0001 on the same voyage moves from 390 to 135 and stays feasible. Recorded with the pricing gate off, the arm in which this expedite is offered as a card; the shipped default carries it as advice.

WHEN SEVERAL CONNECTIONS BREAK AT ONCE

On the cascade pack one inbound slip leaves three connections broken, one AT_RISK and two INFEASIBLE, competing for one shift budget. CP-SAT allocates one action to each in a single solve at $5,600, taken one card at a time. A greedy plan matches the cost on this pack; across the solver study the joint solver saves 423 connections against greedy's 408.

THE SIX BEHAVIOURS IN THE BRIEF, AS CHECKS ON THE LEDGER OF THREE REAL EPISODES (50 EVENTS)

Analyse the input, identify the issue	an ingest event and a fusion event are present
Determine a course of action	feasibility check, options enumerated, rule verdicts recorded
Orchestrate tools and workflows	three or more distinct tools called
Handle uncertainty and tool failure	a completeness verdict, plus a fault, an error or a degraded mode
Invoke human review or escalation	an approval card raised, plus a human decision or an escalation
Produce a clear execution trace	CSA field set on each event, chain verifies, rationale labelled

Steps 1 to 5 and 7 are the graph's ledger on the hero pack; step 6 is the console's what-if flow on the same approval server, margin unchanged at 101. Each check fails on an empty trace.

`scenario_pack_hero.json · cascade-evidence.json · solver_quality.json · behaviours-conformance.json`

This entry adds three mechanisms. Each column names the published pattern it builds on.

1 · THE AGENT PRICES ITS OWN ACTIONS BEFORE PROPOSING THEM

Each candidate action is priced at enumeration inside the twin's replication distribution. An option whose expected value is below its cost is never proposed as a write. It is carried as ADVISE_ONLY with the probability, the value and the cost attached, the solver drops it from the candidate set as it would a refused option, and the connection stays on the board with its margin and no card.

Most expedites do not save a connection. The gate cuts spend per saved connection from $60,835 to $19,745 and turns the rest into advice with the numbers attached.

SAME 500 SCENARIOS	GATE OFF	GATE ON
Expedites executed	173	29
Expedite spend	$138,400	$23,200
Spend per rollover avoided	$60,835	$19,745
At-risk connections carried as advice	0	157 of 299

The gated probability comes from a held-out draw the gate did not select on: 0.0489 in sample, 0.0397 held out. Horvitz's expected-utility threshold (CHI 1999) supplies the rule; the addition here is that the utility is audited on held-out data.

2 · THE OFFICER'S EDIT RE-DERIVES THE POLICY AND RE-BINDS THE TOKEN

An approver can change the plan, and the change cannot widen the agent's authority. The edit resolves to a solver-enumerated option or a priority level, the policy row is re-derived from the edited arguments, and the approval token is re-minted over the edited digest.

The mechanism is extracted as a standalone package. It ships with a refunds workflow as a non-port example and a conformance module that checks the package against RELAY's recorded behaviour.

12 of 12

attacks on the package are refused; two, self-approval and card resurrection, succeeded on the first run and were fixed at the token authority

206 of 206

conformance checks against RELAY pass, 203 of them with byte-identical output

IMDA's framework describes the editable plan; the re-derivation of the policy row and the re-binding of the token are new here.

3 · EACH NAMED CONTROL WAS SWITCHED OFF AND HAD TO MAKE A TEST FAIL

A passing test suite does not by itself show that a named control is wired in. The probe script parses each control the four deliverable documents claim, disables it, and checks that a named test which was green on the clean tree now fails. The denominator is the full set of claimed controls; none was selected out.

56 of 56

probes run were caught; 0 survived, 0 invalid

48 of 56

controls named by the deliverables were probed; 5 are listed unprobed with the reason and 3 live outside this code

Along the way it found a live path traversal that thirteen green tests had missed, and a false reading caused by Python's bytecode cache. Mutation testing is established; applying it to named oversight controls, with the denominator parsed from the documents and a certificate rendered from the run, is new here.

Borrowings come first in the prior-art document: LLM-Modulo, CSA guidance, adaptive sampling, rule-plus-model extraction.

`twin/ev_gate.py · governance/ · evalx/mutation_probes.py · docs/PRIOR-ART-AND-ORIGINALITY.md`

In 500 seeded scenarios the agent flags all 299 at-risk connections; rules alone flag 88.3%, and the 201 clean scenarios produce 0 false escalations.

35

connections the rules cannot see, because the risk exists in the email and in no structured event. All 35 are escalated to the duty officer; none is recovered automatically. The size of that class is a generator parameter, so the figure measures the mechanism.

The sweep is synthetic, and its ground truth is the twin's own feasibility check.

`evalx/results/sweep-full-n500.final.json`

320 of 320

verdicts on which an independent oracle, written from the contract text, agrees with the twin; the rules-only baseline reaches 0.894 against it. Seven single-point mutations of the rule were detected 4 of 7 times; the three missed change no verdict in this scenario set.

The agreement certifies the margin arithmetic and leaves the completeness gate to its own tests.

`evalx/results/validity-oracle-n320.json · oracle-mutation-power.json`

0.726

share of 200 advisories on which the hybrid router extracts every field exactly, against 0.548 for a regex baseline and 0.575 for the local model alone. It flags each seeded contradiction (recall 1.000 against 0.471), invents an ETA in 11.9% of cases against the model's 47.7%, and passes 4 wrong facts end to end against the model's 10.

Its gate routing, 0.850, sits below the baseline's 0.865; the deck reports the lower figure.

`evalx/results/fusion-ladder.json`

12 of 12

prompt-injection advisories produced 0 unsafe tool calls through the real graph. Nine escalate before a tool is chosen, and three reach a tool choice without an unsafe call.

The corpus is 48 adversarial advisories written for this entry, so the figure describes the boundary on this corpus.

`evalx/results/fusion-ladder.json, injection_resistance`

14 of 14

attacks on the approval path are refused after four fixes at the token authority (self-approval, a token spent twice, a race on the single-use spend, a resurrected card). A verifier checks for each write that the executed argument digest equals what the card rendered: 7 checks, 5 of 5 episodes.

Threat model: a compromised agent process that cannot forge the approval server's pepper. Measured with the pricing gate off, where a card exists.

`evalx/results/approval-attacks.json · evalx/oversight_chain.py`

10,000

episodes in a soak run with faults injected into a quarter of them across all ten failure classes: 0 writes without an approval or a policy gate, 0 writes after degradation, 0 chain failures, 0 episodes stuck past 30 seconds.

Single process, replay tier, on the laptop that recorded the demo.

`evalx/results/soak-profile.json`

The write gate and the approval server enforce access control and audit, and a queueing model prices the human oversight.

ACCESS CONTROL

Per-agent credentials for fusion, planner, executor and console; the executor alone may present an approval token to a write tool. An approval token is a SHA-256 over the card, the tool, the argument digest, the approver, the expiry and a server pepper. The approval server is the sole issuer: it mints the token when a card is approved, and the token is spent once under a lock. The deciding principal must be a human account, and the maker of a card cannot be its checker. In this build the approval server runs in-process; the production boundary is a separate service holding the pepper and the token store, which the frozen contract already isolates. The STRIDE-lite review, S-0 to S-24, carries a rerunnable command per row; it found a cross-site request forgery on the console, since closed.

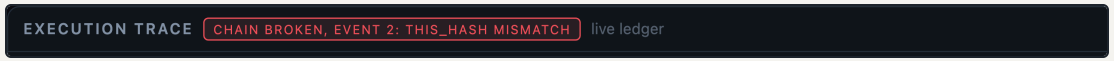
TOKENS AND RUNTIME

Each advisory costs one fusion step: a three-sample panel that escalates to five samples when the cheap panel disagrees. The five-sample build measures 4,689.7 model tokens per advisory episode off the ledger, and adaptive sampling cuts 33.8% on the bench with every quality column unchanged. The local tier runs at $0 imputed. The frontier tier stays off unless an operator sets a key, and a kill switch overrides the routing; at the frontier list price the same model tokens would cost about $0.0026 per advisory episode, or about $317 a day if each of 122,000 daily TEU were its own episode, which is an upper bound.

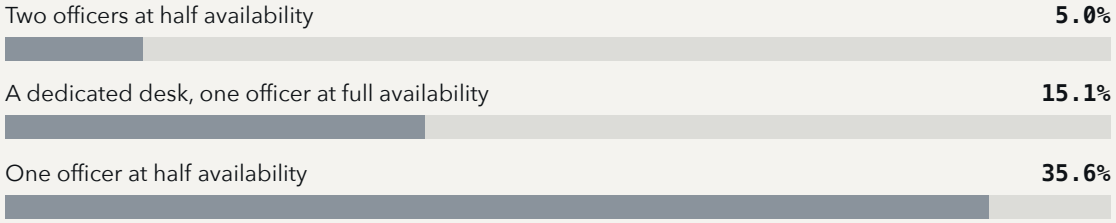
The graph is stateless per episode and checkpointed, so it scales horizontally. Rate counters are shared across processes under a lock, and sixteen concurrent consumers cannot overspend a budget of five. The ledger is the one serialisation point; sharding by correlation id is contracted, and it is not built.

AUDIT AND TAMPER EVIDENCE

Each event sits on one SHA-256 chain with a head anchor, so an edit, a reorder or a truncation is visible and replay refuses a broken chain. Model rationale is stored under its own label and is excluded from the audit record. A soak of 10,000 episodes completed in 21.3 minutes at p99 21.7 ms with faults injected into a quarter of them, with 0 chain failures and 0 invariant failures. The scale profile found a quadratic append, 0.83 ms at chain length 100 and 3.16 ms at 500; after the fix an append costs 0.050 ms at a chain length of 8,000.



MODELLED DENY-BY-DEFAULT RATE BY DESK STAFFING (NO DESK HAS BEEN STAFFED)



Share of approval cards that expire, at one at-risk connection in ten and a 90-second read, from an M/M/c queue with Erlang C checked against a discrete-event simulation over 100,000 cards, on the pre-gate card rate of 7.9 cards an hour as an upper bound. Mitigations not yet built: a per-class window in the policy table, a second approver before a denial, cards batched per inbound vessel.

The ledger is tamper-evident: an adversary with root access and the source can rewrite it, so production places the store outside the agent's credential scope.

The gate turns a $2.4M loss into +$87,215 a year for the chain, positive with probability 0.683; PSA's column is negative because a rolled box counts as billable storage.

ANNUAL RESULT, CHAIN-WIDE, BASE YEAR

Value of kept connections, before operations	+$0.3M
of which shipper +$1.2M, carrier +$0.2M, PSA −$1.1M	
Desk and supervisor time	−$243,328
Net of operations, shipped gate	+$87,215
PSA's own column, net of operations	−$1.3M
First build, gate off, rejected	−$2.4M

Each of the model's 43 inputs carries one of four labels: MEASURED from a results file, CITED with a source, CHOSEN with a range, or GENERATOR-DERIVED from a named simulator constant. The volume basis is PSA's 44.5 million TEU of 2025 with the transhipment share and rollover chain the model labels; on it the shipped gate produces 20,883 escalations a year, about 0.066 of one officer's reading time.

WHY THE PSA COLUMN IS NEGATIVE

A saved connection is a box that does not sit in the yard. In the model PSA then forgoes storage of $20 per box-day that it would have billed and gains a yard slot worth $5 per box-day. Both are CHOSEN inputs with no source found, and the storage charge is the assumption to replace first, because a transhipment box inside its free time is not billed.

At the most favourable corner of every stated range at once, storage waived among them, a rollover avoided is worth $7,200 to PSA against $27,152 to the chain at base inputs, itself a modelled value built from CHOSEN inputs, and an $800 expedite still does not pay PSA. PSA's line reaches zero only if a freed yard slot were worth $125.58 per box-day, against the model's range of $0 to $15. Re-priced on PSA's column the gate proposes no writes, which is the gate working as intended on a commercial arrangement that sends the value elsewhere.

THE DEPLOYMENT CASE

The value of a kept connection lands with the shipper and the carrier. A case for PSA therefore rests on a commercial change: the expedite funded by the party that receives the value, or the recovery priced into the Assured Port Time product. One unquantified item is stated as such: a kept connection is transhipment volume retained at a hub that sells reliability.

The sign of the chain-wide figure rests on 0.0397 against a break-even of 0.0370. A bootstrap of the 29 held-out values puts the probability between 0.0293 and 0.0509, and the annual figure is positive with probability 0.683. The largest drivers are the price of a rebooking proposal, a CHOSEN input, and boxes per connection, GENERATOR-DERIVED; connection outcomes from PSA would replace both with measurements.

A kept connection avoids the late re-route and the wasted yard move inside PSA's Scope 3 Category 9, its largest category at 347.2 ktCO2e. This deck claims the mechanism and does not estimate a tonnage saving.

The terminal data is synthetic and the limits are listed. Each request to PSA would replace one assumption with a measurement.

DATA SOURCES AND ASSUMPTIONS

A synthetic terminal world, labelled SYNTHETIC in the data itself. The project holds no PSA data and no API keys, and it is not integrated with PORTNET; the adapter is stubbed against a frozen contract.

Recorded Singapore AIS, 24 and 25 August 2026: 151,906 messages, 1,551 vessels, used as pseudonymised aggregates.

PSA Annual Report and Sustainability Report 2025 for the cited figures: 84% Assured Port Time, 44.5 million TEU, Scope 3.

Sea-Intelligence schedule reliability of 62.6% in June 2026 for the probability that a vessel is late. Port of Barcelona berth allocation instances, downloaded at run time under GPL-3, for the external solver benchmark in the appendix.

The advisory lead time of 30 to 240 minutes is a CHOSEN value and is labelled as one wherever it appears.

SEE IT RUNNING

PROJECT PAGE

thylinao1.github.io/relay-page

EXPLANATORY VIDEO, 9 MINUTES

youtu.be/J3wU0hluJfU

AUDIO PODCAST, 22 MINUTES

thylinao1.github.io/relay-page/#listen

LIMITS

The 500-scenario sweep is simulator-internal, graded by the twin's own feasibility check under a simulated approver who never declines. The independent oracle reduces that objection and leaves it open. The generator gives the structured stream the fact 85% of the time, the conservative setting, which understates the model's contribution; the recorded AIS days show the opposite, and both figures are published.

A rebooking is a proposal until the carrier grants it; the margin does not move before that, and the countdown board labels it pending.

No desk has been staffed with this system, so the 35.6% expiry rate comes from a queueing model.

The local 3B model's parse layer invents an ETA in 47.7% of the cases where none exists; the deterministic layer contains that, and the hybrid router brings it down to 11.9%.

Refusal as a solver constraint changes no plan under the shipped gate (59 of 59 agree); with the gate off it is better on 26 of 60 scenarios and worse on 0.

Access control is at demo scope, and token expiry is bound to the world clock so that replays are deterministic.

REPRODUCE

Contract stubs and fixtures, 18 checks	**ALL PASS**
Full test suite	**1,351 passed**
Oracle gate, 28 hand checks	**reproduced**
Scripted demo, headless	**all steps pass**

Four commands in the README, no network; the test suite is about eight minutes of it. The demo video accompanies this deck; code, tests and every result file are available to the judges on request.

WHAT RELAY NEEDS FROM PSA FOR THE SPRINTERNSHIP

Anonymised connection outcomes, to calibrate the twin and replace the CHOSEN rows in the impact model, the storage charge first.

Interface documentation for PORTNET and the Service Allocation Tool, so RELAY runs between planning cycles.

PSA's own alert taxonomy, so the tiers line up with the categories the remote operations team already uses.

The first deployment is advisory: a shadow-mode run beside the duty desk, with cards shown and not executed, measured by officer acceptance rate and time to decision. T1 writes follow that evidence.