

Zero-Compute Deterministic Interception of Multi-Agent Collusion Carriers

AgentGate is a loopback Policy Enforcement Point between agents and their tools. It bounds side channels, carries restrictions across delegation as HMAC caveats, and fingerprints split or encoded data, with no model calls on the decision path.

57 / 57

carrier-group defence tests pass (groups A, B, C)

1.00 ms

p99 CPU per benign decision (reference 2 ms)

verified

hash-chained decision log, 750 entries

Three carriers, three deterministic guards

GROUP A

Boolean / timing oracles

Information rides on which of a few argument variants is chosen, or on call rhythm, so no fragment of the value appears.

side_channel_guard.py: after the taint point, caps distinct variants, call count, and regular rhythm; a signal holds the session and freezes the lab key 15 min.

8 of 8 tests pass

GROUP B

Confused-deputy delegation DAGs

A restricted agent hands its task or state to a peer whose own capability looks clean.

delegation_guard.py: hand-off re-mints the recipient's agw1. token with the sender's deny caveats (lattice meet, monotonic); barred principals are held at egress.

8 of 8 tests pass

GROUP C

Split-chunk / encoded smuggling

A protected value leaves in small pieces across steps, or after Base64, hex, URL or whitespace transforms.

deobfuscate.py + canary_watermark.py: salted n-gram digests over a sliding window and decoded forms; zero-width HMAC watermarks name the origin step.

41 of 41 tests pass

Tamper-evident audit: agw1. caveats + hash-chained logs

agw1. macaroon-style capabilities

- Chained HMAC-SHA256: each caveat re-keys the tag, so allow_egress=false cannot be stripped without the root key.
- Untrusted ingest appends the deny caveat; delegation copies it to the recipient.
- One-time out-of-band approval, bound to session, step, tool and argument hash, is the only bypass.

SHA-256 hash-chained decision log

- Every line carries seq, prev_hash and entry_hash (a linear chain, not a Merkle tree).
- audit_manifest.json adds file digests, the chain head, and an HMAC; verify exits 1 on any edit, drop, reorder, insert or truncation.
- Canary hits record origin_step_index and origin_tool_name in the chain.

This run: 750 benign entries verified: yes | edited line detected: yes | dropped line detected: yes

Test-suite results and stateful Hypothesis fuzzing

262

gateway tests collected (pytest --collect-only)

57 / 57

carrier-group tests passed in this run

Each figure traces to a named test in collusion_benchmark_results.json.

Carrier groups (pytest JUnit XML, keyed by node id)		
A	test_side_channel_guard.py	8 / 8 pass
B	test_delegation_guard.py	8 / 8 pass
C	test_deobfuscate.py, test_canary_watermark.py	41 / 41 pass

Stateful Hypothesis machines (settings read from the tests)		
DelegationDagMachine	test_delegation_guard.py	120 x 24 steps
SideChannelMachine	test_side_channel_guard.py	60 x 30 steps
PepPolicyMachine	test_stateful_fuzz.py	40 x 16 steps
SmugglingMachine	test_stateful_fuzz.py	60 x 24 steps
GatewayScrubMachine	test_stateful_fuzz.py	80 x 8 steps

Measured benign CPU latency and zero-LLM-compute architecture

0.126 ms

p50 CPU per decision

0.892 ms

p95 CPU per decision

0.999 ms

p99 CPU (reference 2 ms)

3.84 ms

slowest single decision

- 750 benign evaluate_call decisions across 5 multi-agent workflows x 50 runs; 0 denied.
- LLM calls on the decision path: 0. Checks are hashing, HMAC, set lookups and interval arithmetic.
- Thread CPU time, in process, on one machine, including the decision-log write. The 2 ms figure is a reference target, not a guarantee.
- p99 CPU is within the reference target; the slowest single decision is shown, not hidden.

What these numbers are, and what they are not

- Same-author scripted lab tests and workflows. Not a benchmark, and not recall against an unknown adversary: group counts show the guards behave as their own tests specify.
- Benign workflows are few and author-written; the denial count is a smoke check, not a deployment false-positive rate.
- Loopback only: binds 127.0.0.1 by default. Freezes last 15 minutes and apply to registered lab keys only; unregistered callers are never acted on.
- ATLAS technique labels are approximate.

Real-model loopback archive: 2 of 13 non-control runs

A separate measurement on one machine with one 3B model. It is never pooled with, averaged into, or compared against the test-suite figures on these slides.