Deploy the Hermes Agent connector

Use this guide to deploy one aip-host-hermes-agent instance for a fixed set of tenant-owned Hermes endpoints. The procedure proves host and endpoint readiness, introduces one protected read, and replaces or rolls back replicas without movin

SECTION	Connectors
TYPE	Connector
PROTOCOL	AIP 1.0
SOURCE	docs/connectors/hermes-agent/operations/deployment.md

Complete the common connector fleet deployment first. This page adds the Hermes endpoint, operator, MCP, and rollout decisions. The endpoint configuration and standalone host are distinct from the Hermes API-server deployment itself.

Preserve one endpoint-set boundary

One connector instance owns one tenant and one immutable endpoint set. Each endpoint contributes 35 endpoint-qualified capabilities to the manifest. Applications call product-neutral `aipd`; the fleet gateway routes an admitted Action to this host's `/aip/v1/messages` endpoint.

Every overlapping replica for the instance must use:

- the same tenant, membership, external-account boundary, endpoint set, and

admitted connector version;

- the same durable PostgreSQL state;
- a distinct pre-provisioned replica ID and signing seed;
- endpoint credentials accepted by the same reviewed credential policy.

The endpoint file accepts up to 1,000 descriptors, but the common host rejects more than 512 manifest capabilities. At 35 capabilities per endpoint, this revision admits at most 14 endpoints in one standalone host. Treat 14 as a manifest ceiling, not a throughput qualification.

1. Freeze the deployment record

Record the following identities before starting the host:

Identity	Hermes deployment rule
Connector type and version	Match the admitted Hermes manifest and immutable artifact
Instance ID	Represents one tenant-owned endpoint set
Replica ID	Identifies one concrete process; never reuse it concurrently
Tenant and membership	Match every endpoint descriptor and fleet binding
External account	Name the stable Hermes deployment or endpoint set, normally with system `hermes_agent`
Public endpoint	End in `/aip/v1/messages` and resolve to this replica
Host DID and seed	Match the pre-provisioned replica identity

Identity	Hermes deployment rule
Artifact digest	Identify the exact running bytes
Manifest digest	Bind the normalized endpoint IDs and all 35-operation groups
Configuration revision	Increase when deployment-owned configuration changes
Credential revision	Identify the revision allowed for new route assignments
Pinned Hermes revision	Identify the provider route and event contract reviewed for this release

Changing an endpoint ID changes capability IDs. Changing an endpoint URL or tenant changes routing or ownership. Admit either change as a reviewed manifest and configuration update rather than editing a live replica in place.

2. Prepare durable state and network paths

Provision one durable PostgreSQL boundary for the connector instance. The common host stores Actions, results, approvals, streams, callbacks, and recovery state there. The Hermes host also attaches its operator bindings and delegated-result resolver to the same runtime stores.

Allow only the required flows:

Direction	Peer and route	Purpose
Inbound	Fleet gateway to `/aip/v1/messages`	Signed manifest-pinned native AIP traffic
Inbound, operator network	`/health`, `/ready`, `/metrics`, `/aip/v1/manifest`	Process and host observation
Outbound	Connector lifecycle control endpoint	Signed registration, heartbeat, drain, and offline transitions
Outbound	Connector PostgreSQL	Durable runtime and Hermes operator state
Outbound	Every configured Hermes `base_url`	Fixed provider API routes
Outbound, optional	Central callback endpoint	Signed stream and result delivery

Use TLS and reviewed service discovery across trust boundaries. The shared connector-host CA config protects host, control-plane, and callback clients; the Hermes connector's provider client uses the process or container trust store instead. Install any private provider CA there explicitly.

Do not expose observation routes to an untrusted network. Do not route application traffic directly to the connector host or Hermes API server.

3. Mount endpoint configuration and secrets

Set the shared values from Connector host configuration, then add the required endpoint file and optional operator-policy file:

```
TEXT
AIP_HERMES_ENDPOINTS_FILE=/etc/aip/hermes-endpoints.json
AIP_HERMES_OPERATOR_POLICY_FILE=/etc/aip/hermes-operator-policy.json
```

The endpoint file is a strict JSON array. A one-endpoint deployment can use:

```
JSON
[
  {
    "id": "primary",
    "base_url": "https://hermes.internal.example",
    "api_key_file": "/run/secrets/hermes-primary-token",
    "display_name": "Primary Hermes",
    "tenant_id": "tenant-acme"
  }
]
```

The process rejects an empty array, unknown descriptor fields, unsafe URLs, normalized ID collisions, tenant mismatch, and unsafe secret files. Endpoint API keys are read once at startup. Replace the process after rotating one.

Mount these values as owner-controlled files:

- PostgreSQL URL and the 32-byte Ed25519 signing seed;
- each endpoint API key;
- gateway and lifecycle DIDs when file-based configuration is selected;
- the optional private CA and credential-revision policy;
- the endpoint and optional operator-policy JSON documents.

Secret readers reject symlinks, non-regular files, empty or oversized values, and group- or world-accessible Unix permissions. Keep token bytes out of the endpoint JSON, environment, image, logs, and admission package.

Omitting the operator-policy file keeps first-class delegation disabled and uses the reviewed built-in operator policy. Do not enable delegation until both capability and scope allowlists, durable approval verification, and the inside-Hermes AIP MCP path have been reviewed independently.

4. Admit the exact artifact and endpoint set

Build and publish `aip-host-hermes-agent` through the release process, then pin the image or binary by immutable digest. Verify that the manifest contains exactly 35 capabilities for every normalized endpoint ID and no unexpected endpoint.

Pre-provision the connector instance and each replica in the registry. The host may register only its existing replica identity; it does not create its own registry object. Compare these values before admission:

- type, version, instance, replica, tenant, and membership;
- public endpoint, host DID, gateway DID, and trust domain;
- artifact, manifest, and schema-bundle digests;
- endpoint count and capability IDs;
- region, zone, capacity class, config revision, and credential revision.

Keep tenant bindings for chat, runs, sessions, jobs, operator execution, and delegation disabled until a read-only canary has passed.

5. Start and prove readiness

The reviewed host performs this sequence:

1. parse shared and Hermes-specific arguments;
2. read, normalize, and tenant-bind every endpoint descriptor;
3. read endpoint keys and the optional operator policy;
4. build and bind the immutable manifest to the host signer;
5. open PostgreSQL and recover durable runtime state;
6. attach operator profile state, approvals, Action queue, and lifecycle store;
7. probe storage and `GET /health` on every Hermes endpoint;
8. register the exact pre-provisioned replica and acquire its lease;
9. start heartbeat renewal and serve the native host routes.

`/health` proves only that the host process can answer. Require `/ready` to report all four conditions:

- the host is not draining;
- the registry lease is current;
- durable runtime storage is ready;
- connector health is ready.

Hermes connector health calls unauthenticated `GET /health` on every configured endpoint. It proves that those routes were reachable within the health bound. It does not prove endpoint Bearer authentication, model-provider readiness, MCP connectivity, protected routes, mutations, or operator execution.

Compare `/aip/v1/manifest` with the admitted digest. Confirm that the registry shows the expected replica, endpoint, DID, topology, lease, and no unexpected active assignment before enabling discovery.

6. Introduce one protected read

Enable only the intended tenant binding and one endpoint's low-risk discovery group. Through the authenticated fleet gateway:

1. discover `cap:hermes_agent:<endpoint_id>:models` for the intended tenant;
2. verify that another tenant cannot discover it;
3. invoke it with a fresh Action ID;
4. retain the route assignment, result, receipts, and redacted host evidence;
5. confirm that the response came from the intended endpoint and credential.

The protected model-list call proves more than the readiness probe because it uses the endpoint Bearer token. It still does not qualify chat, streaming, runs, sessions, jobs, tools, or a configured model route.

Expand one capability family at a time. Review its approval, idempotency, timeout, data, and recovery contract before enabling mutations. Add AIP MCP inside Hermes and first-class delegation as separate changes with independent principals and evidence.

7. Replace a replica without moving ownership

For a normal replacement:

1. admit the new immutable artifact and a new replica identity;
2. start it against the same PostgreSQL and endpoint-set boundary;
3. require manifest parity, registration, a current lease, and `/ready`;
4. route only the reviewed canary scope under fleet policy;
5. compare protected reads, error rates, latency, leases, and endpoint evidence;
6. reconcile every uncertain, active, approval-pending, or replica-pinned Action;
7. terminate the old process with `SIGTERM` or `Ctrl-C`;
8. require its drain and registry `offline` transition;
9. expand the replacement only after the retained evidence passes review.

Normal shutdown sets draining before asking the registry to drain. The host rejects new work, waits for in-flight requests up to its drain timeout, stops serving and heartbeats, and requests the offline transition. Set the platform termination grace above that timeout plus network margin.

Do not run two replicas with the same replica ID or signing seed. Do not point overlapping replicas at independent runtime databases.

Roll back without replaying uncertain work

Disable the affected binding in a higher signed package revision before restoring an earlier version. New discovery and assignments must stop, but provider work already accepted by Hermes may continue.

Reconcile Actions against the replica and durable state that own their route. Drain the rejected version normally and retain its PostgreSQL state, operator bindings, event history, callback records, logs, and lifecycle evidence. Restore the earlier artifact together with its matching manifest, endpoint file, operator policy, and admission record.

Never retry an uncertain chat, run, session mutation, job mutation, operator start, approval command, or stop through another replica merely because the rollback finished.

Resolve deployment failures

Symptom	Safe decision
Process exits before registration	Validate strict endpoint JSON, file bounds, tenant equality, URLs, identities, and complete grouped settings
Manifest exceeds the host limit	Reduce the endpoint set; do not raise a nonexistent Hermes-specific ceiling
Registration is rejected	Compare every admitted identity, digest, endpoint, revision, and topology value
`/health` succeeds and `/ready` fails	Separate lease, drain, PostgreSQL, all-endpoint health, and registration evidence
One endpoint keeps the host unready	Remove it through a reviewed manifest change or restore its health; do not hide the failed probe
Protected read returns `401` or `403`	Replace the correct endpoint key and process; do not put a token in Action input
Private provider CA is untrusted	Install it in the process trust store; the shared host CA setting does not configure this client
Old replica accepts no new work	Check whether it is intentionally draining before restarting it
Old replica never becomes offline	Keep new mutations paused and follow the host lifecycle recovery procedure
Operator state is missing after replacement	Stop rollout and verify that both replicas use the same durable runtime stores
Delegated result integrity fails	Preserve both records and fail closed; do not accept Hermes prose as a result

Retain deployment evidence

Retain one record with:

- AIP, connector-host, and pinned Hermes revisions;
- artifact, manifest, schema, endpoint-file, and operator-policy digests;
- instance, replica, tenant, membership, endpoint, and external-account IDs;
- host and peer DIDs, database identity, topology, and network policy;
- configuration and credential revisions without secret values;
- readiness response, registry lease, protected canary result, rollout decision,

drain result, and offline state.

This evidence applies only to the recorded artifact, endpoint set, credentials, tenant, and time window. It is not an all-capability or live-provider qualification.

Related documentation

- Deploy the connector fleet
- Connector host configuration
- Connector host lifecycle
- Hermes connector configuration
- Hermes authentication and endpoints