

Metrics reference

AIP exposes process health, work capacity, fleet state, and bounded activity counters through three different metric surfaces. This reference is for operators who need to select the correct endpoint, recognize every current product-neutral

SECTION	API Reference
TYPE	Reference
PROTOCOL	AIP 1.0
SOURCE	docs/reference/metrics.md

This inventory describes source revision `97be86e9efedf07ecf1783b03800f683f107fb04`. It covers `aipd`, the shared host used by Cal.diy, Hermes Agent, Chatwoot, Dify, CrewAI, and Twenty, and the connector lifecycle control plane. It does not invent provider-specific business metrics.

Surface selector

Process	Path	Format	Current inventory
<code>aipd</code>	<code>GET /metrics</code>	Prometheus text, version 0.0.4	37 possible families: 34 direct daemon plus 3 embedded gateway
Connector host	<code>GET /metrics</code>	Prometheus text, version 0.0.4	8 fixed families
Connector control plane	<code>GET /metrics</code>	JSON	6 connection-pool fields

The three paths have the same name but not the same representation. Do not configure a Prometheus text parser against the control-plane JSON endpoint.

Stable names versus current emission

The observability crate declares 11 stable metric names for AIP implementations. A declaration fixes a shared name and meaning; it does not register or emit a series.

Stable name	Declared meaning	Emitted at this revision
<code>aip_envelopes_received_total</code>	Envelopes received by a gateway	Yes, by the embedded gateway
<code>aip_envelopes_sent_total</code>	Envelopes sent by a gateway	No
<code>aip_actions_total</code>	Action invocations	Yes, by the embedded gateway
<code>aip_delegations_total</code>	Delegation requests	No
<code>aip_delegation_duration_ms</code>	Delegation latency in milliseconds	No
<code>aip_action_duration_ms</code>	Action latency in milliseconds	No
<code>aip_connector_invocations_total</code>	Connector invocations	No
<code>aip_connector_duration_ms</code>	Connector latency in milliseconds	No
<code>aip_errors_total</code>	Protocol errors	No
<code>aip_sessions_active</code>	Active sessions	No
<code>aip_stream_chunks_total</code>	Emitted stream chunks	No

Only the two rows marked **Yes** are emitted by the pinned product-neutral processes. Do not create alerts or dashboards for the other nine until an implementation registers them. No product-neutral histogram is emitted at this revision.

aipd Prometheus families

Readiness and native traffic

Metric	Type	Updated by	Meaning
<code>aip_daemon_ready</code>	Gauge	Daemon <code>/ready</code> request	Composite daemon readiness: gateway, worker, required NATS listener, and required modules
<code>aip_gateway_ready</code>	Gauge	Gateway readiness evaluation	Gateway storage and registered connector readiness
<code>aip_runtime_worker_up</code>	Gauge	Worker lifecycle and daemon <code>/ready</code>	Whether the durable background worker is running
<code>aip_nats_listener_up</code>	Gauge	NATS listener lifecycle and daemon <code>/ready</code>	Whether the native NATS listener is running; readiness requires it only when configured as required
<code>aip_envelopes_received_total</code>	Counter	Gateway envelope entry	Envelope attempts received before policy completion, labeled by <code>message_type</code>
<code>aip_actions_total</code>	Counter	Gateway envelope entry	Received envelopes whose body is an action, before policy completion

`aip_actions_total` is not a success counter. Pair it with durable action state and error evidence when calculating outcomes.

Fleet maintenance and registry pools

Metric	Type	Updated by	Meaning
<code>aip_connector_fleet_ready_replicas</code>	Gauge	Successful fleet maintenance cycle	Ready replicas in the current fleet summary
<code>aip_connector_fleet_active_assignments</code>	Gauge	Successful fleet maintenance cycle	Current active route assignments
<code>aip_connector_fleet_expired_replicas_total</code>	Counter	Successful fleet maintenance cycle	Replicas expired by maintenance since this process started
<code>aip_connector_fleet_maintenance_failures_total</code>	Counter	Failed fleet maintenance cycle	Expiry or fleet-status cycle failures since process start
<code>aip_connector_registry_control_pool_connections</code>	Gauge	Successful cycle with a pool snapshot	Open control-plane database connections
<code>aip_connector_registry_control_pool_idle_connections</code>	Gauge	Successful cycle with a pool snapshot	Idle control-plane database connections
<code>aip_connector_registry_control_pool_max_connections</code>	Gauge	Successful cycle with a pool snapshot	Configured control-plane connection ceiling
<code>aip_connector_registry_data_pool_connections</code>	Gauge	Successful cycle with a pool snapshot	Open data-plane database connections
<code>aip_connector_registry_data_pool_idle_connections</code>	Gauge	Successful cycle with a pool snapshot	Idle data-plane database connections
<code>aip_connector_registry_data_pool_max_connections</code>	Gauge	Successful cycle with a pool snapshot	Configured data-plane connection ceiling

The six pool gauges describe the registry owned by `aipd`; they do not describe the separate lifecycle control-plane process.

Remote admission and event ingress

The remote admission gauges are refreshed immediately before `aipd` renders a metrics scrape, but only when the scheduler is configured.

Metric	Type	Meaning
<code>aip_connector_remote_active</code>	Gauge	Active admitted remote actions
<code>aip_connector_remote_queued</code>	Gauge	Queued remote actions
<code>aip_connector_remote_queued_bytes</code>	Gauge	Serialized bytes retained in the admission queue
<code>aip_connector_remote_active_tenants</code>	Gauge	Tenant partitions with active remote work
<code>aip_connector_remote_queued_tenants</code>	Gauge	Tenant partitions with queued remote work

Event-ingress capacity gauges are also refreshed at scrape time when central connector event ingress is configured.

Metric	Type	Meaning
<code>aip_connector_event_ingress_capacity</code>	Gauge	Configured maximum concurrent ingress requests
<code>aip_connector_event_ingress_available</code>	Gauge	Currently available ingress permits
<code>aip_connector_event_ingress_received_total</code>	Counter	Requests reaching either connector-events or connector-callbacks handler
<code>aip_connector_event_ingress_accepted_total</code>	Counter	Ingress envelopes accepted by the shared handler
<code>aip_connector_event_ingress_rejected_total</code>	Counter	Requests rejected for missing configuration, schema, decode, trust, route, capacity, or durable handling failure

Both ingress URL paths use the same counters. The metrics do not separate connector events from stream callbacks and do not carry connector, replica, or tenant labels.

Runtime work budgets and retention

Metric	Type	Updated by	Meaning
<code>aip_runtime_callback_capacity</code>	Gauge	Worker cycle	Maximum concurrent callback-delivery work
<code>aip_runtime_callback_available</code>	Gauge	Worker cycle	Available callback-delivery permits
<code>aip_runtime_reconciliation_capacity</code>	Gauge	Worker cycle	Maximum concurrent reconciliation work
<code>aip_runtime_reconciliation_available</code>	Gauge	Worker cycle	Available reconciliation permits
<code>aip_runtime_worker_cycles_total</code>	Counter	Completed worker cycle	Worker cycles completed since process start
<code>aip_runtime_worker_claimed_actions_total</code>	Counter	Completed worker cycle	Durable queued actions claimed by this process
<code>aip_retention_events_deleted_total</code>	Counter	Hourly retention pass	Event records deleted since process start
<code>aip_retention_replay_claims_deleted_total</code>	Counter	Hourly retention pass	Replay claims deleted since process start
<code>aip_retention_dead_letters_deleted_total</code>	Counter	Hourly retention pass	Dead-letter records deleted since process start
<code>aip_retention_callback_deliveries_deleted_total</code>	Counter	Hourly retention pass	Callback-delivery records deleted since process start
<code>aip_retention_outbox_records_deleted_total</code>	Counter	Hourly retention pass	Outbox records deleted since process start

The capacity gauges express concurrency permits, not queue depth. A value of zero available with nonzero capacity means the corresponding work class is saturated at that sample.

Connector-host Prometheus families

Every one of the six standalone connector hosts renders all eight families on each scrape. The endpoint uses local atomics and has no metric labels.

Metric	Type	Meaning
<code>aip_connector_host_ready</code>	Gauge	<code>1</code> only when the host is not draining, its lease is valid, storage is ready, and the connector is ready
<code>aip_connector_host_storage_ready</code>	Gauge	Last known durable runtime-storage readiness
<code>aip_connector_host_in_flight</code>	Gauge	Action requests currently executing inside this host
<code>aip_connector_host_requests_total</code>	Counter	Native message requests received by this process
<code>aip_connector_host_rejected_total</code>	Counter	Requests rejected by the instrumented schema, trust, route, readiness, capacity, credential, approval, or gateway paths
<code>aip_connector_host_heartbeat_failures_total</code>	Counter	Lease renewal or recovery attempts that failed from the heartbeat loop
<code>aip_connector_host_lease_recovery_attempts_total</code>	Counter	Attempts to re-register after a lease is no longer valid
<code>aip_connector_host_lease_recoveries_total</code>	Counter	Lease recoveries that successfully applied a new lease

`requests_total - rejected_total` is not a supported success calculation. Response-signing and process failures can occur outside a counter update, and accepted action state is owned by the durable lifecycle rather than this arithmetic.

Control-plane JSON snapshot

The connector lifecycle control plane returns a `ConnectorRegistryPoolSnapshot` as JSON:

```
JSON
{
  "control_size": 1,
  "control_idle": 1,
  "control_max": 4,
  "data_size": 1,
  "data_idle": 1,
  "data_max": 4
}
```

Field	Meaning
<code>control_size</code>	Open registry administrator/control connections
<code>control_idle</code>	Idle control connections
<code>control_max</code>	Configured control connection ceiling
<code>data_size</code>	Open lifecycle data connections
<code>data_idle</code>	Idle lifecycle data connections
<code>data_max</code>	Configured data connection ceiling

These values are a point-in-time JSON diagnostic. A Prometheus collector must transform them explicitly or use a separate JSON exporter; they are not native Prometheus series.

Labels and cardinality

The current product-neutral Prometheus emitters use one metric label: `message_type` on `aip_envelopes_received_total`. Its value comes from the bounded AIP message-type enum. Every other current daemon and host family is unlabeled.

Stable tracing fields such as action ID, message ID, session ID, correlation ID, delegation ID, connector ID, and error code are not metric labels. Keep those values in traces and structured logs. Do not add tenant, action, resource, provider request, free-form message, or replica identifiers as Prometheus labels.

Use target metadata outside the application to distinguish processes, regions, connector types, and immutable deployment artifacts. Keep the target-label set bounded and controlled by the monitoring system.

Series activation and reset behavior

Prometheus recorder families are registered lazily when code updates them. Consequently:

- `aip_daemon_ready` and `aip_gateway_ready` may be absent until readiness has been evaluated;
- remote admission and ingress gauges are absent when their optional subsystem is not configured;
- fleet and pool gauges are absent until a successful maintenance cycle, and pool gauges also require a pool snapshot;
- traffic, failure, and retention counters may be absent until their first update;
- the nine stable names with no emitter remain absent for the entire process.

Absence is not zero. Monitor the scrape target and the corresponding health or configuration boundary separately. All counters and gauges are process-local; they reset when the process restarts. Use rate or increase semantics for counters and tolerate a reset.

Operational interpretation

Signal	Combine with	Interpretation
Daemon ready is 0	<code>/ready</code> body and required-module state	One or more readiness prerequisites failed
Runtime worker up is 0	Worker error and durable queue state	Background recovery, callbacks, retention, and queued execution may not progress
NATS listener up is 0	Deployment configuration	Incident only when NATS is required for this deployment
Ready replicas fall	Fleet maintenance status, leases, and host readiness	Routing capacity is shrinking; do not infer a provider outage from this gauge alone
Active assignments remain high while ready replicas fall	Drain and lease state	Existing pinned work may be preventing safe removal
Available admission or ingress permits reach 0	Queue gauges and rejection rate	A bounded work class is saturated
Event rejected counter rises	Structured error code and ingress logs	Diagnose schema, trust, route, capacity, or storage; the counter has no reason label
Heartbeat failures rise	Host readiness, lease expiry, and control-plane health	Lease renewal or recovery is failing
Recovery attempts rise without recoveries	Control-plane error and immutable registration	Recovery cannot obtain and apply a valid lease
Retention counters stop changing	Worker cycles and current retained volume	May be normal when nothing is eligible; inspect durable state before alerting

Deployment-owned service objectives determine alert thresholds and windows. This source defines signals and semantics, not universal thresholds.

Endpoint security

At this revision, each metrics path is mounted outside the process's native application-auth middleware. Keep it on a private operations network or place it behind authenticated, authorized, and encrypted monitoring ingress.

The current families deliberately avoid tenant and resource identifiers, but they still reveal capacity, readiness, connection-pool, traffic, and failure information. Do not expose them as public diagnostics.

Related operations

- [Observe and recover](#)
- [Operator runbooks](#)
- [Errors and retry decisions](#)
- [Native HTTP API](#)
- [Connector fleet HTTP API](#)