

[>_]

bulletproof-a2a-gateway – System Overview

Authenticated, rate-limited, and
audited capability invocation.

The Core Mission: A Secure Capability Front Door



Govern

Authenticates every caller via API keys and enforces strict per-caller rate limits.

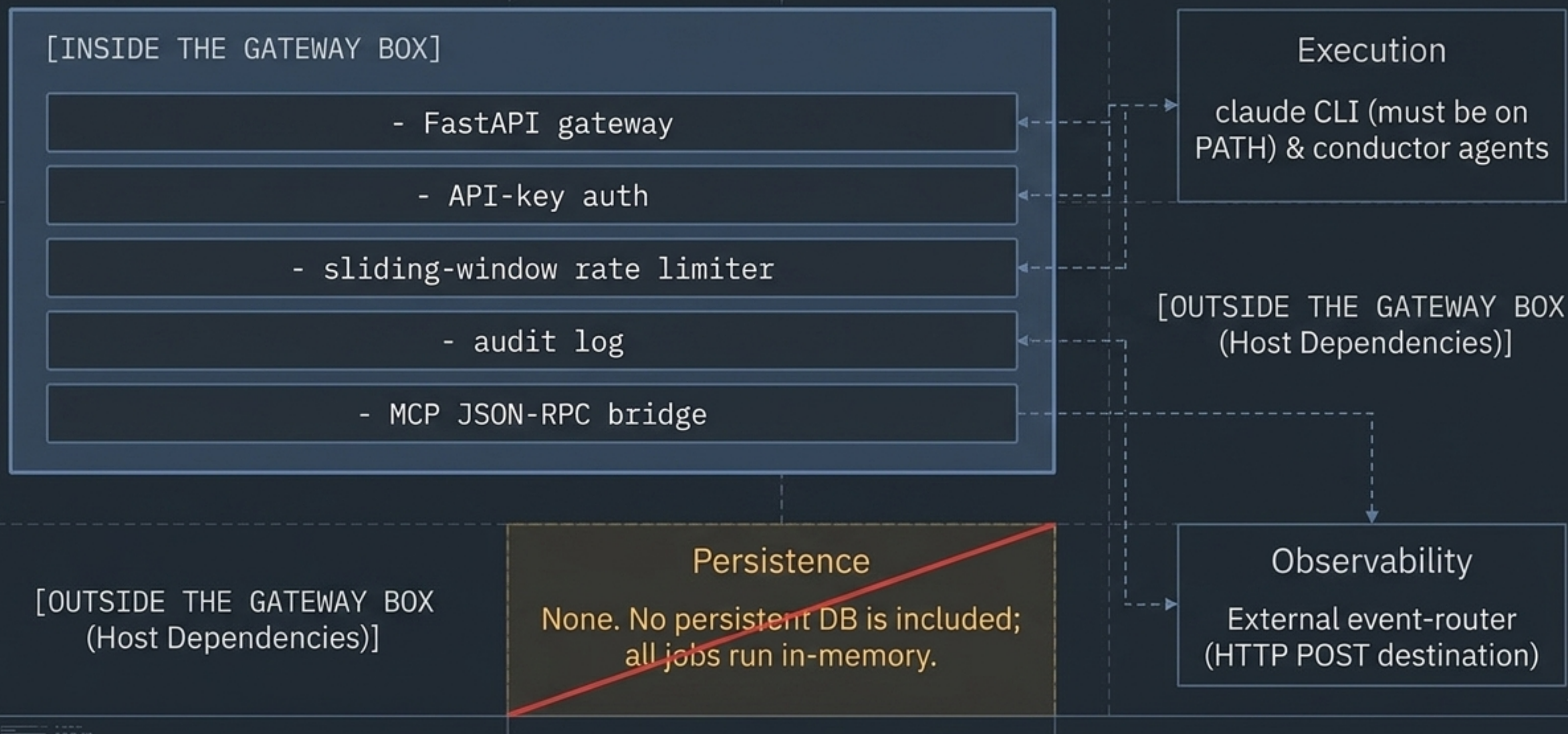
Execute

Shells out to local capabilities safely without bundling the agents themselves.

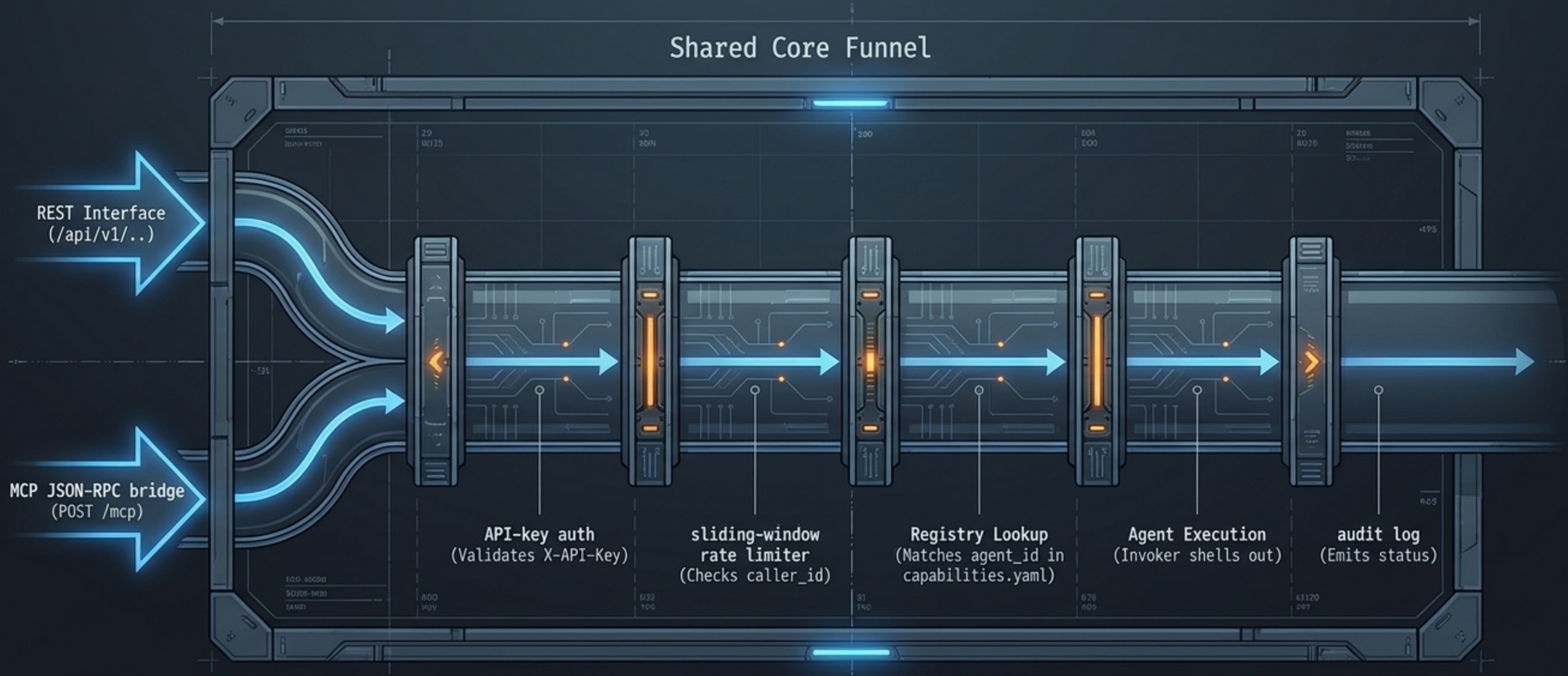
Bridge

Translates standard REST requests and Model Context Protocol (MCP) interactions into a unified pipeline.

Defining the System Boundaries



Architecture at a Glance: Two Surfaces, One Core



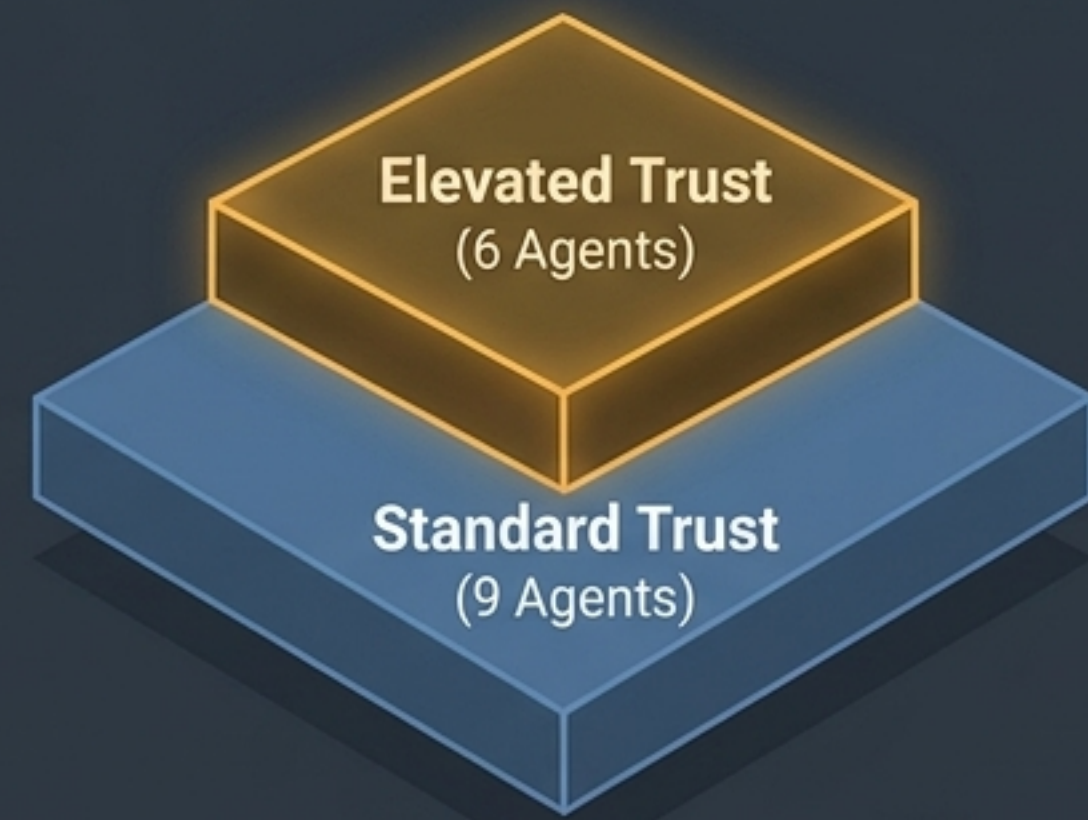
Authentication & Trust Level Boundaries

Primary Authentication



- Driven by API-key auth via the `X-API-Key` header.
- Configured via the `API\_KEYS` environment variable.
- No per-key identity mapping—any valid key passes the gate.

Trust Levels (Registry Defined)



- Invocable by any caller presenting a valid API key.
- Requires an additional `X-Trust-Level: elevated` header. Includes: `architect`, `builder`, `refactor`, `database`, `devops`, `compliance`.

Traffic Control: Sliding-Window Rate Limiter



Mechanism

In-memory sliding-window rate limiter applied after successful authentication.

Parameters

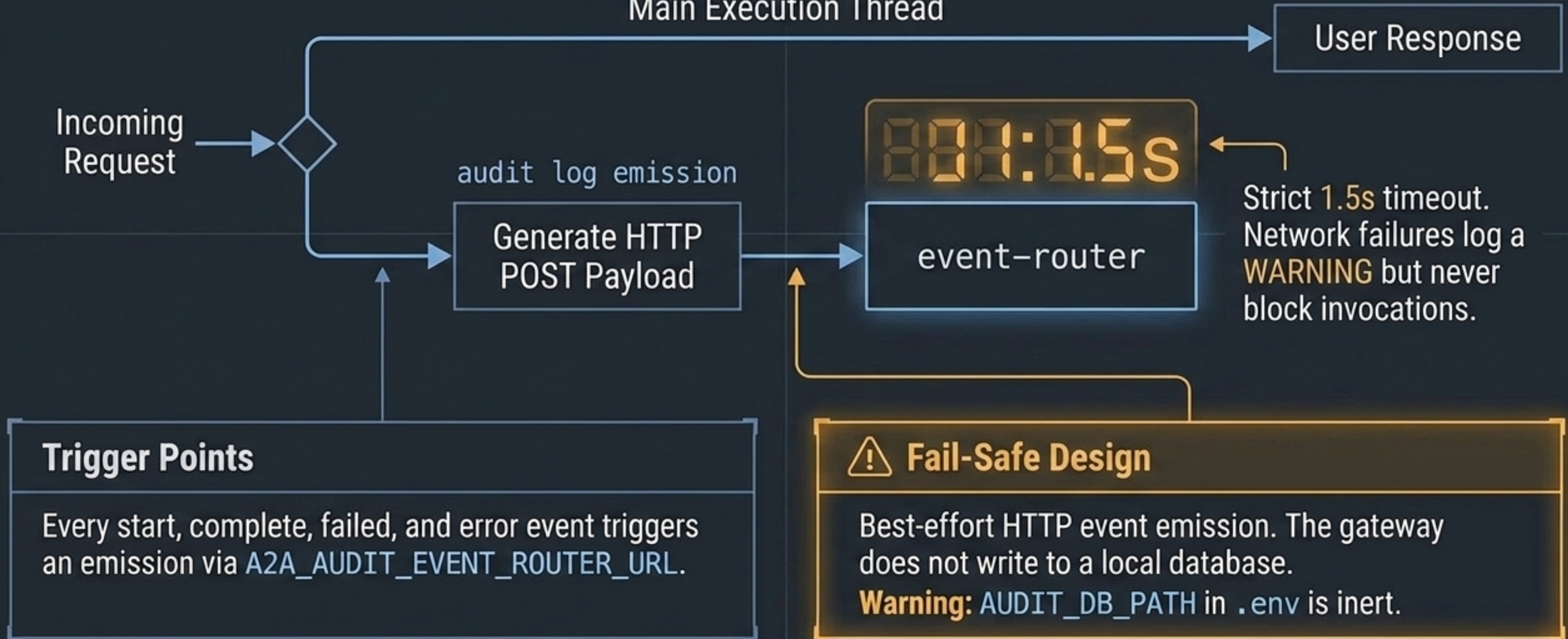
- Hard-coded limit: 60 requests per 60 seconds.
- Tracking key: ``caller_id`` (Shared ``caller_id``'s share the same bucket).

⚠ Operational Scaling Realities

- State is strictly process-local (per-replice).
- Behind a load balancer, effective limits scale linearly with replica count.
- **Warning:** ``RATE_LIMIT_RPM`` environment variable is currently inert; limits are hardcoded.

Observability: The Audit Log Pipeline

Main Execution Thread



The Execution Engine: Registry & Invoker

● ● ● registry/capabilities.yaml

```
>_  
externally_callable: true  
trust_level: standard  
max_tokens: 16384
```

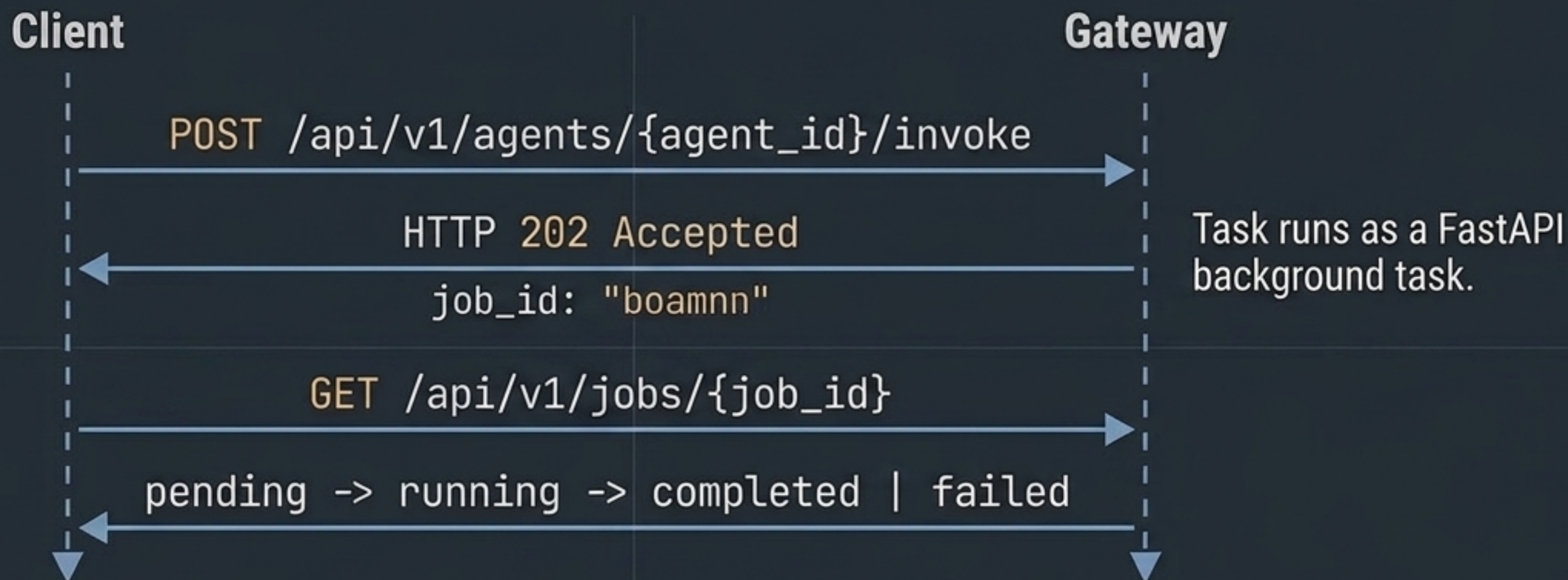
Loaded once at import time. Only
externally_callable agents are exposed.

! The Invocation Subprocess

```
claude -p "<prompt>"  
--agent conductor:<name>
```

- Gateway strips the 'conductor-' prefix from the agent_id.
- Spawns an async subprocess.
- Enforces a strict '300-second' timeout, returning a failed job/error if exceeded.

Entry Surface 1: Asynchronous REST API



State Constraints

Jobs live in a **process-local memory dict**. They **do not survive restarts** and cannot be shared across replicas. Jobs across or have cannot be shared across replicas.

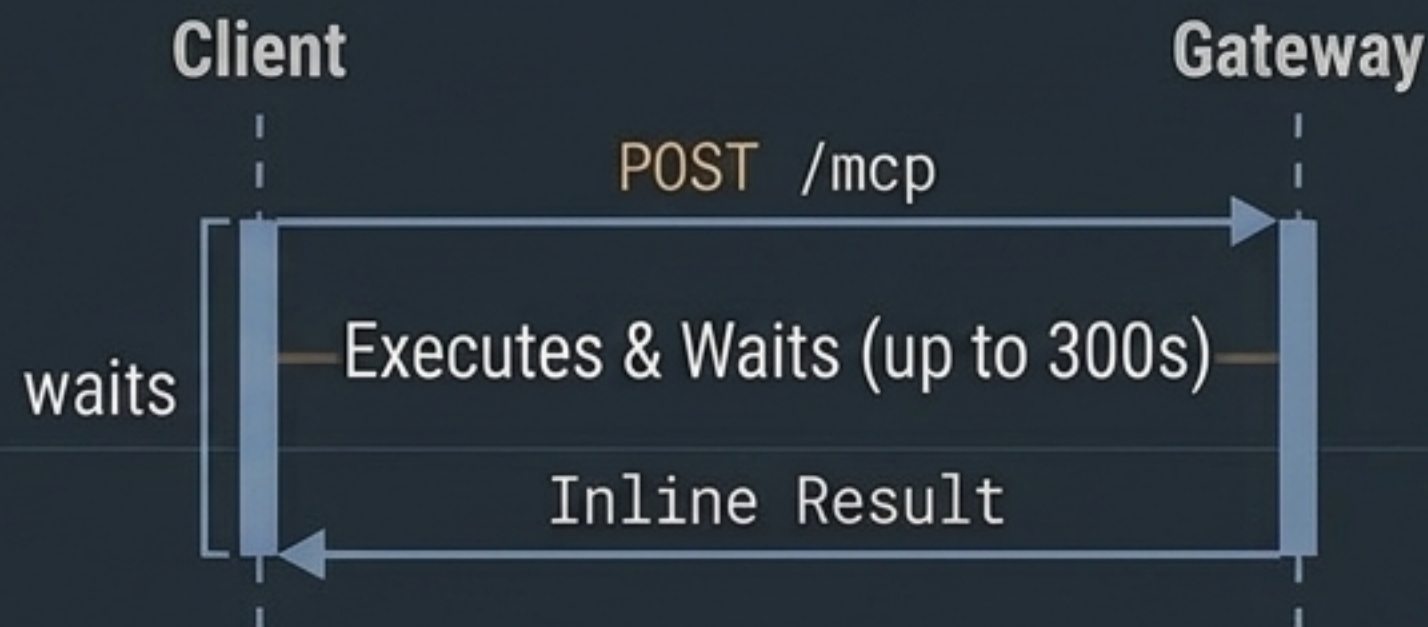
Entry Surface 2: MCP JSON-RPC Bridge

JSON-RPC Protocol

POST /mcp

- Supported methods:
- initialize
 - tools/list (Exposes the 15 conductor agents as distinct MCP tools)
 - tools/call

Synchronous Execution Paradigm



Unlike REST, execution blocks and waits for the agent, returning the result inline (**up to 300s timeout**).

Unified Safety

Utilizes the exact same **API-key auth** and **sliding-window rate limiter** pipeline.
caller_id defaults to 'mcp-client' if omitted.

Surface Comparison: REST vs. MCP

	REST API	MCP Bridge
Protocol & Format	HTTP REST / JSON	HTTP POST / JSON-RPC 2.0
Execution Model	Asynchronous (Returns 202, requires polling job_id).	Synchronous (Waits up to 300s, returns inline).
Trust Level Enforcement	Known Gap — Elevated check is currently a stub; not enforced by the application.	Strictly Enforced — Returns JSON-RPC -32603 on missing X-Trust-Level header.

Security Posture & Hardening

Vulnerability Scans



0 Critical / 0 High vulnerabilities.
Code Hardener score: 942/1000.

Container Identity



Non-root execution.
Docker image drops privileges and runs strictly as appuser (uid 10001).

CI/CD Pipeline



Least-privilege GitHub Actions workflow utilizing SHA-pinned mutable action tags to prevent supply chain poisoning.

Transport Security



The gateway speaks plain HTTP by design. Terminal TLS must be configured at the reverse proxy/ingress layer (Bring your own transport security).

Operational Realities & Known Gaps



CRITICAL SYSTEMS ALERT - CONFIGURATION & COMPLIANCE

Inert Configuration Variables RATE_LIMIT_RPM and AUDIT_DB_PATH are present in .env but are inert in the code. Rate limits are hard-coded; audit is HTTP-only.

REST Elevated Trust Gap Access must be restricted at the network layer for REST if elevated trust enforcement is required, as the code check is currently stubbed.

Port Mismatch .env and README specify 8100, while Dockerfile and registry default to 8420. Operators must explicitly set the port.

Hardcoded Health /health endpoint version is hardcoded to '1.0.0' while the application is actually v1.1.0.

Synthesis: A Precision Front Door

“

The bulletproof-a2a-gateway effectively centralizes governance, observability, and safety, allowing distinct ecosystems (REST architectures and MCP clients) to leverage standard capabilities without exposing underlying agent execution environments.

”

