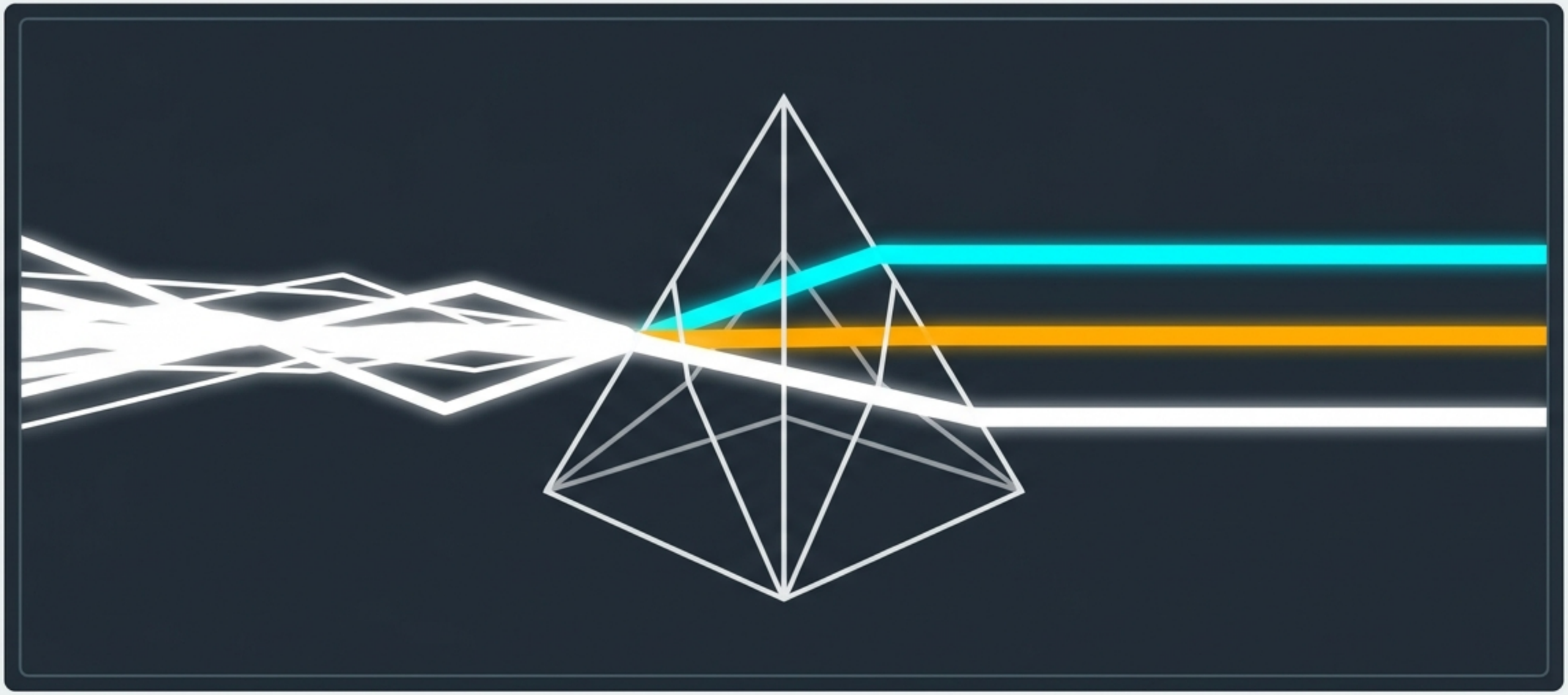


Scaling MCP to Production

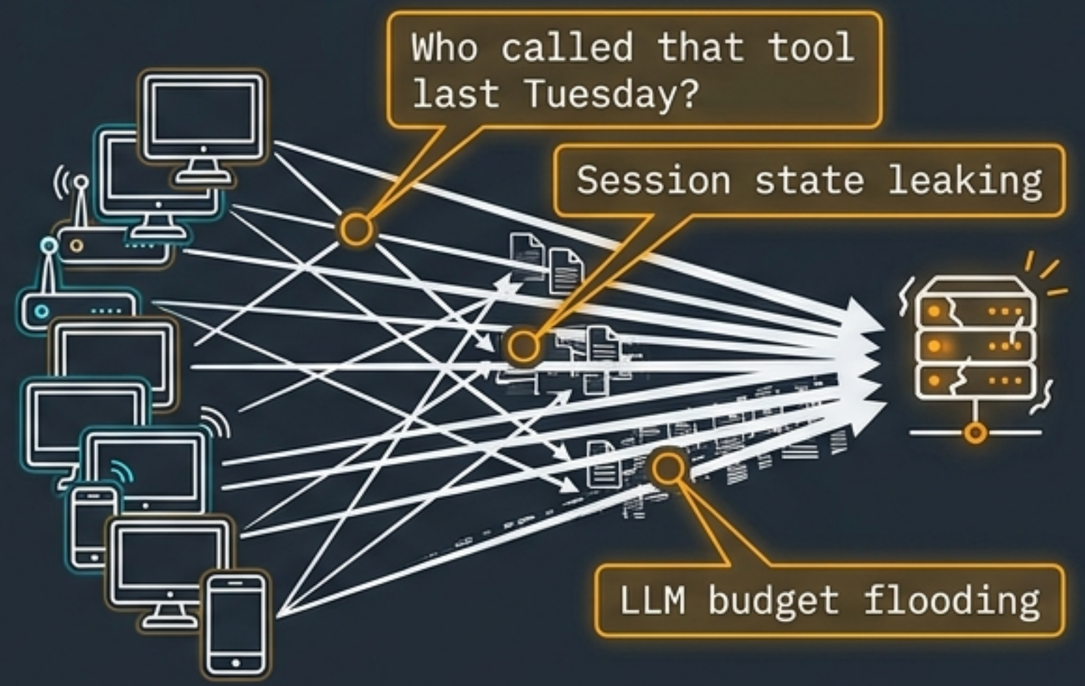
Gateways, Audit Logs, and the Day-1 Playbook for 1,000 Users

The Gap: Local Prototype vs. Enterprise Reality

The Local Prototype

Works on my machine.
No auth, no logs, one user.

The Enterprise Reality



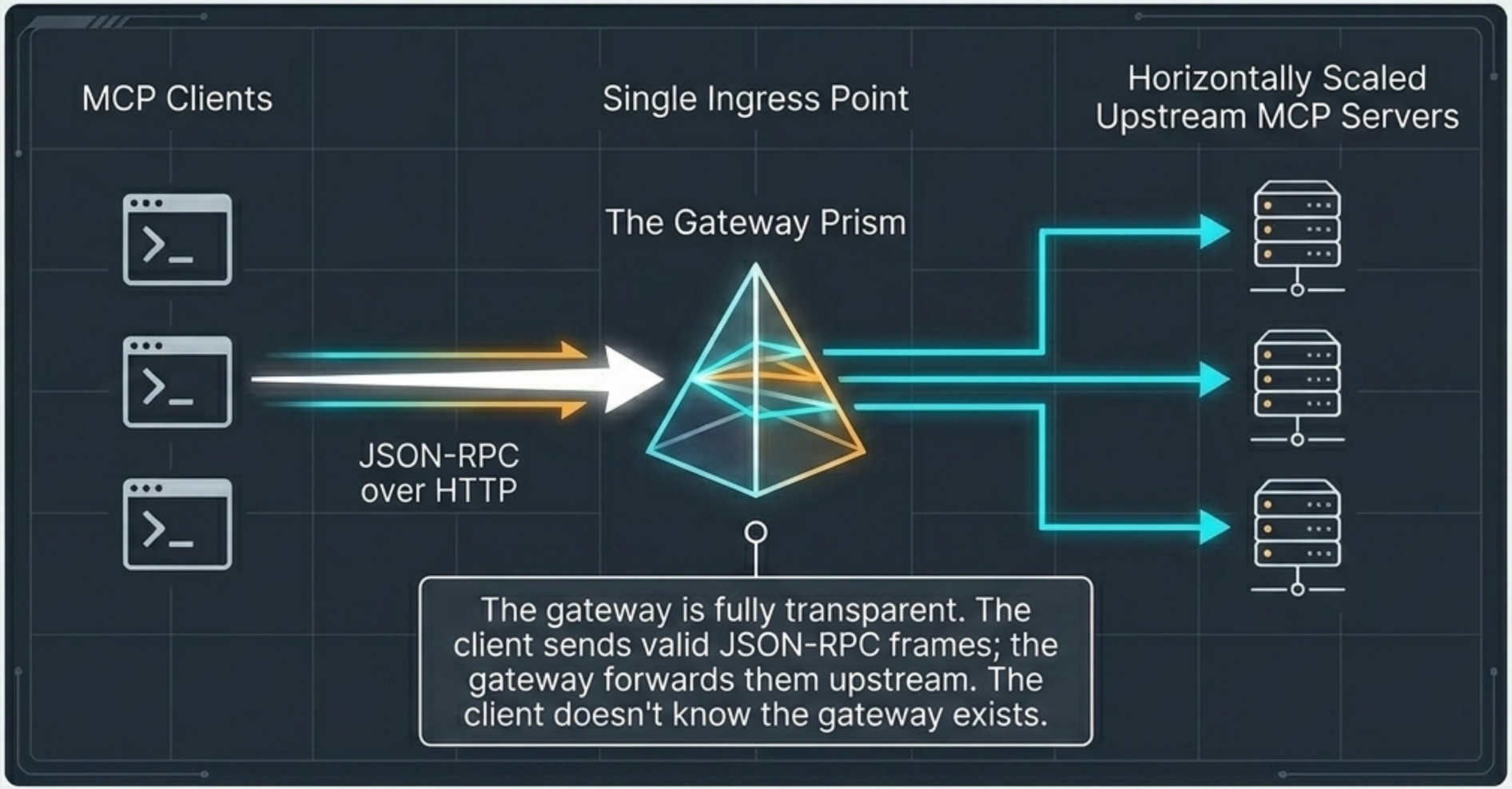
The gap between local development and **multi-tenant production** is entirely **operational**: **security, visibility**, and **scale.**

The Economics of the Day-1 Gateway

	Day 1 Gateway	Day 100 Retrofit
Implementation Time	Under 2 hours	Weeks of refactoring server logic
Audit Readiness	JSON Lines SOC 2 stream immediately	Audit trails lost forever
Architecture	Transparent proxying (zero changes to client/server)	Tangled, bolted-on security inside individual servers

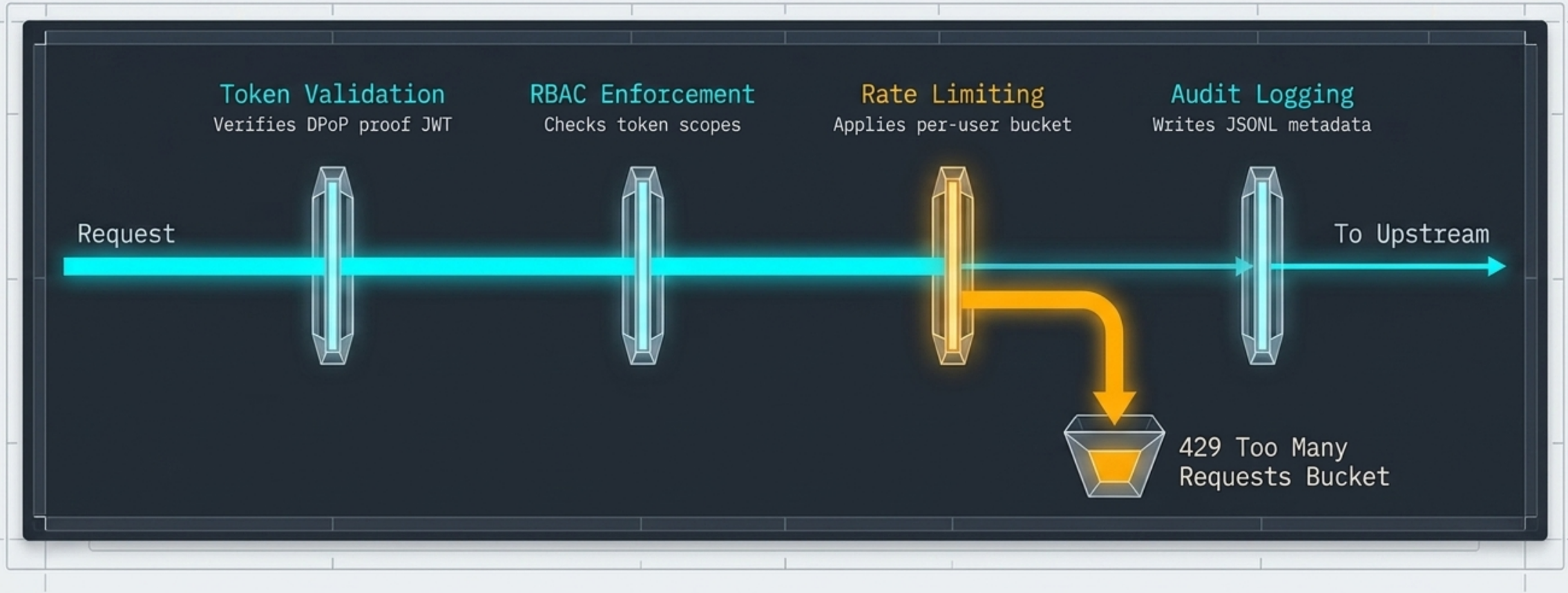
Adding a gateway on Day 1 is an economic choice, not an architectural one.

The Production Topology

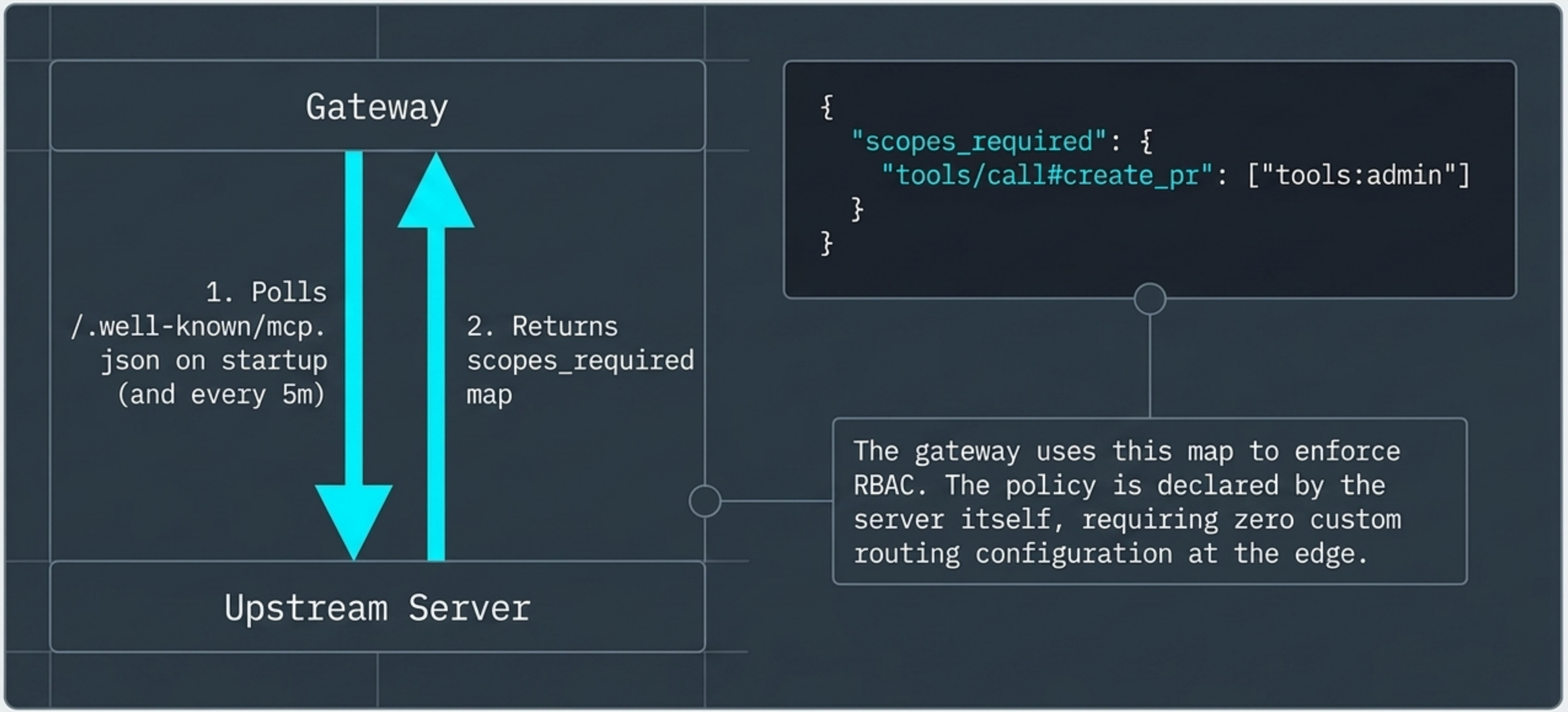


The gateway architecture enables seamless security, observability, and scale without altering existing client-server interactions.

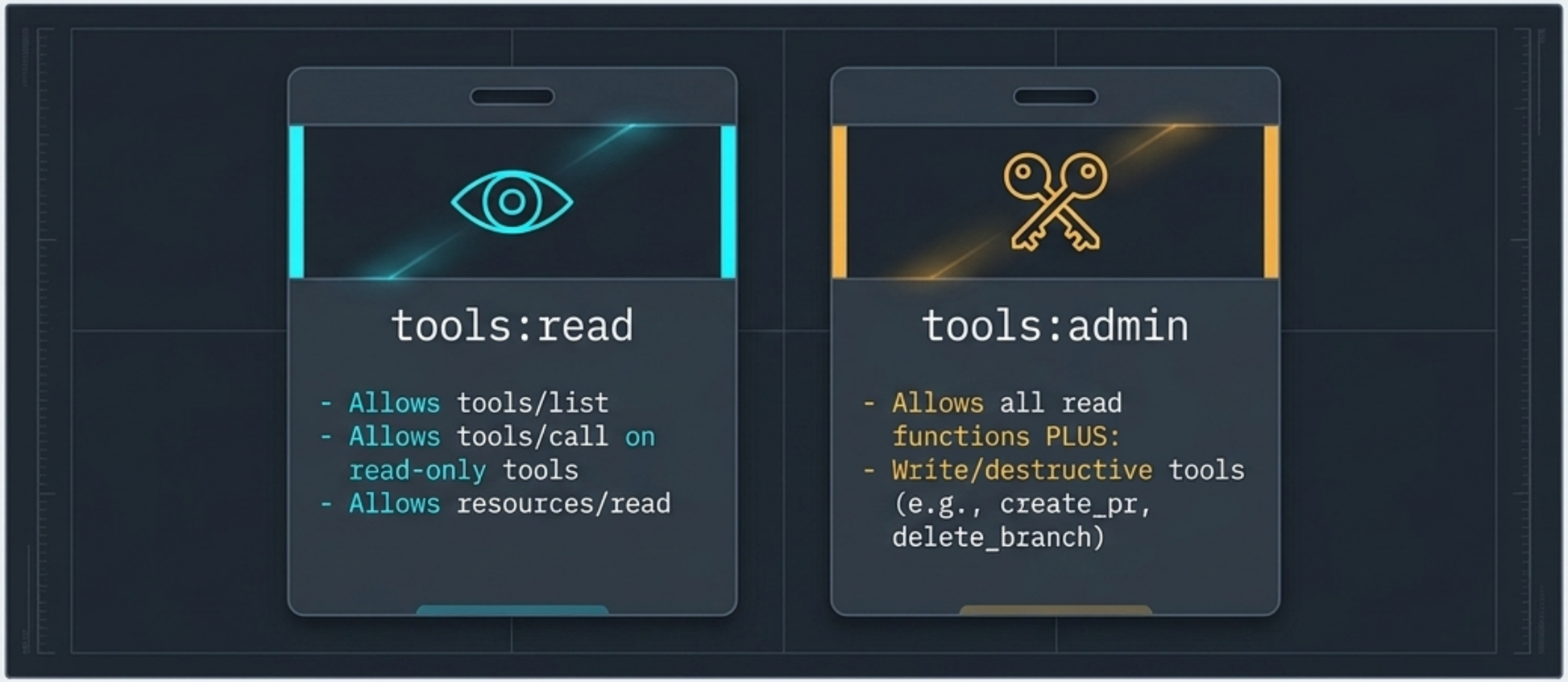
The Edge Request Pipeline



Zero-Touch Configuration via Server Discovery



Scope-Based RBAC & Least Privilege

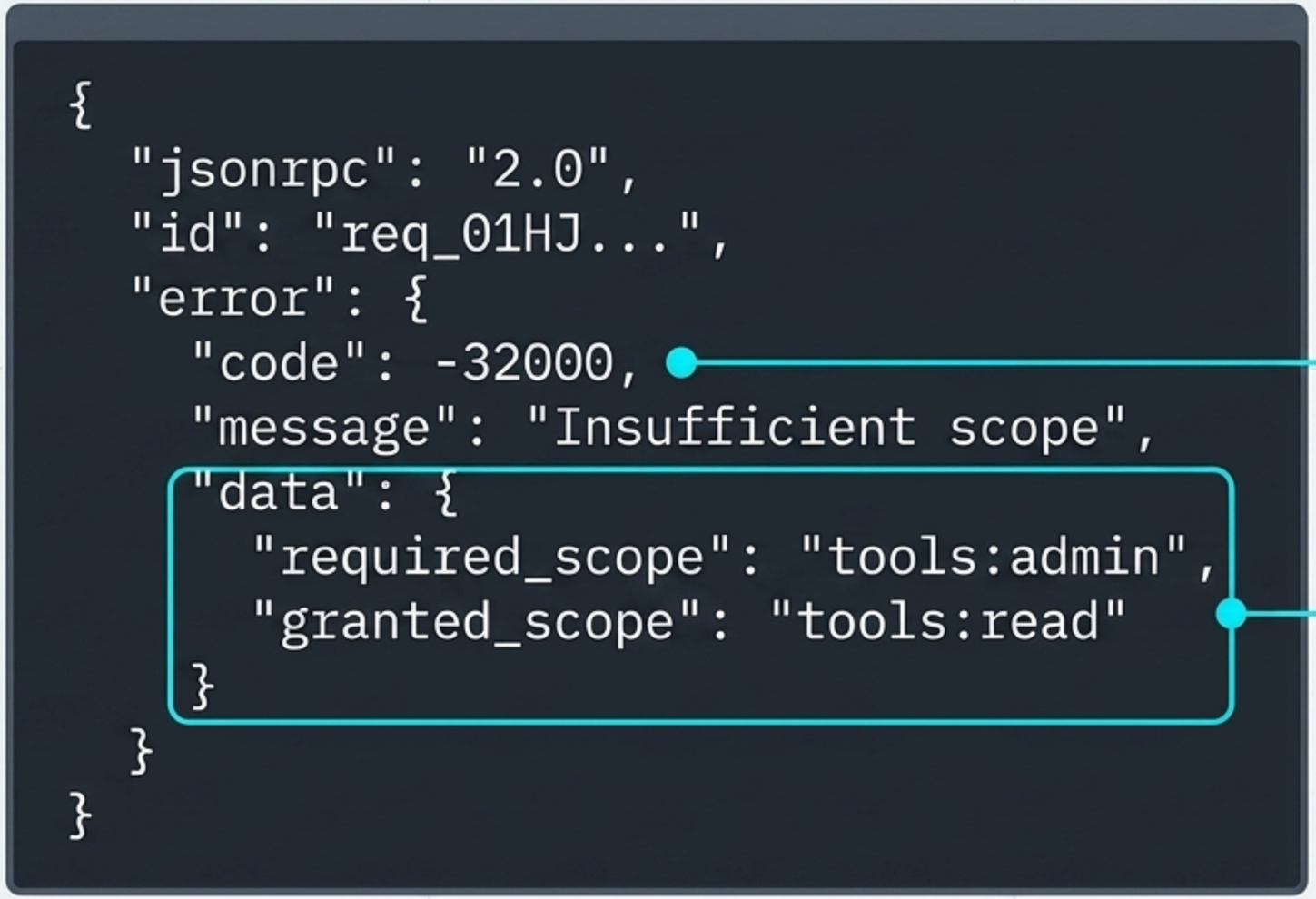


Least-Privilege Principle: Grant the minimum scope needed. A CI/CD pipeline reading file contents should never hold `tools:admin`.

The Access Mapping Matrix

	list_repos (Read)	read_file (Read)	create_pr (Admin)	delete_branch (Admin)
Developer (tools:read)	✓	✓	403	403
Senior Dev / Ops (tools:read, tools:admin)	✓	✓	✓	✓
CI/CD Service Account (tools:read)	✓	✓	403	403
Security Auditor (tools:read - audit view only)	✓	✓	403	403

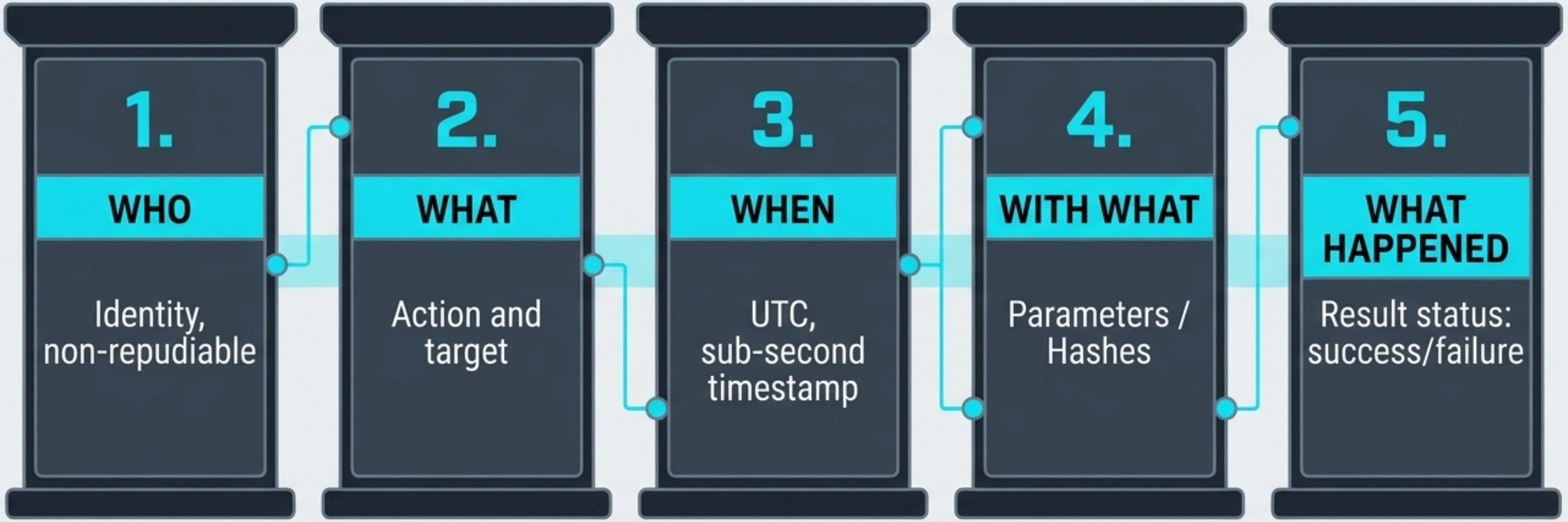
Anatomy of an RBAC Rejection



Standard JSON-RPC application error range.

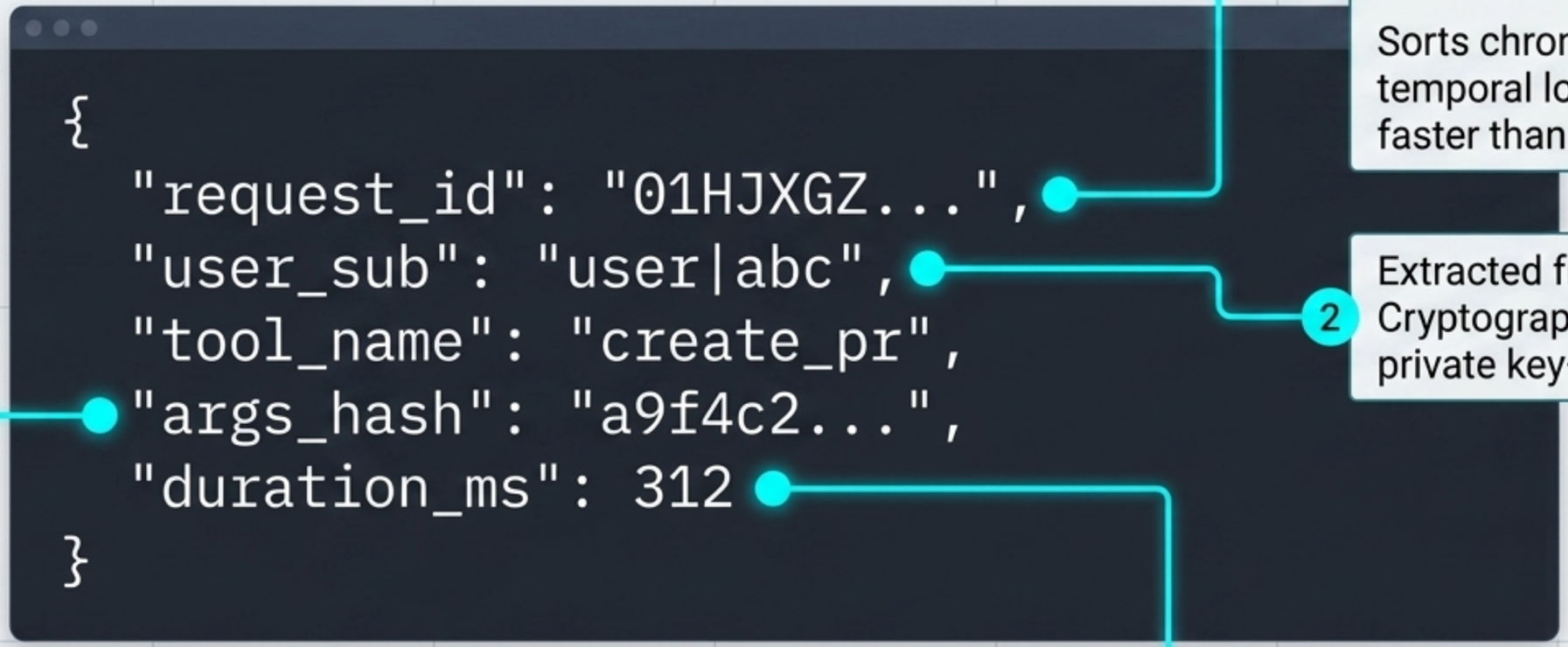
Machine-readable context allows the MCP client to intelligently prompt the user to request elevated access, rather than failing silently.

The SOC 2 Type II Mandate



If your system cannot definitively answer these five questions for every tool call, you cannot pass a security audit.

The Gateway Log X-Ray



3 A gateway-generated ULID.

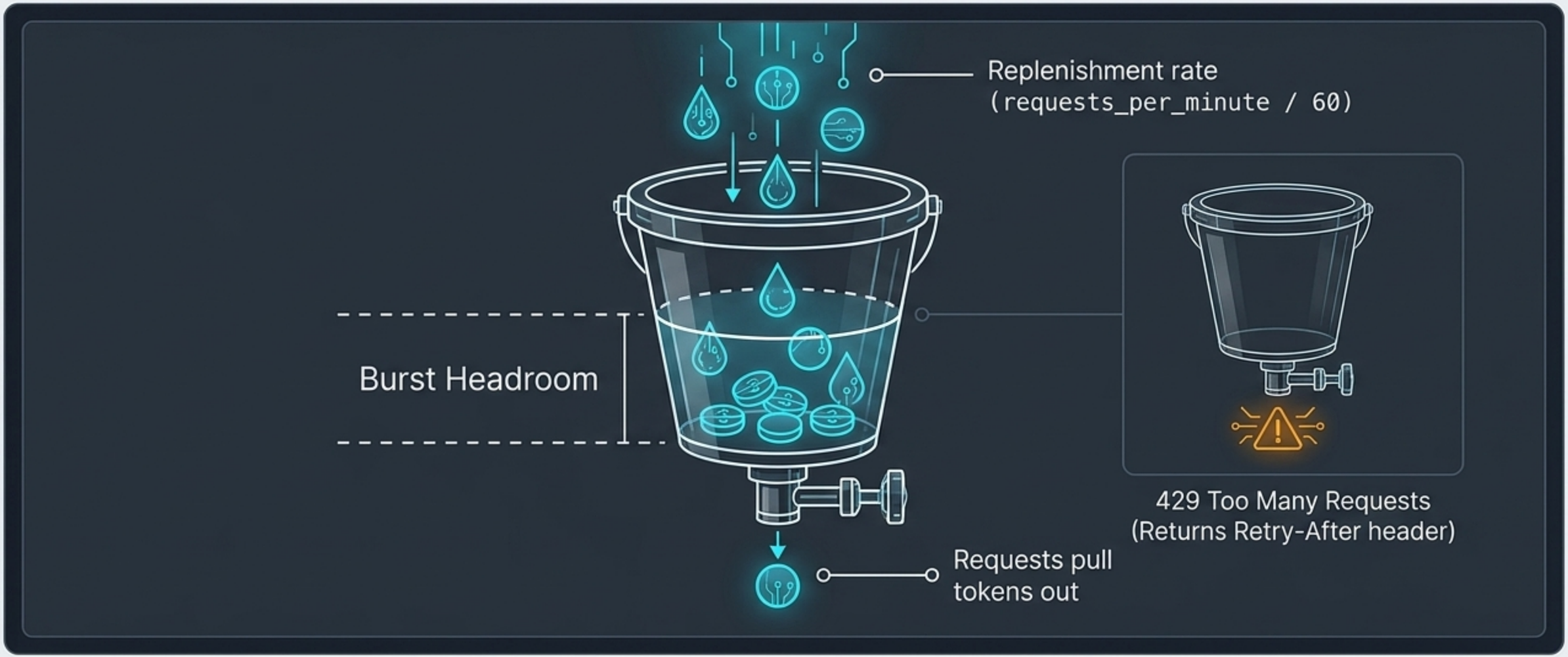
Sorts chronologically, making temporal log queries dramatically faster than UUIDs.

2 Extracted from the DPoP token. Cryptographically tied to the client's private key—strictly non-repudiable.

1 Never log raw tool arguments containing PII or secrets. SHA-256 hash allows audit correlation without data exposure.

4 Gateway-to-upstream round-trip only. Distinguishes slow tool execution from slow client networks.

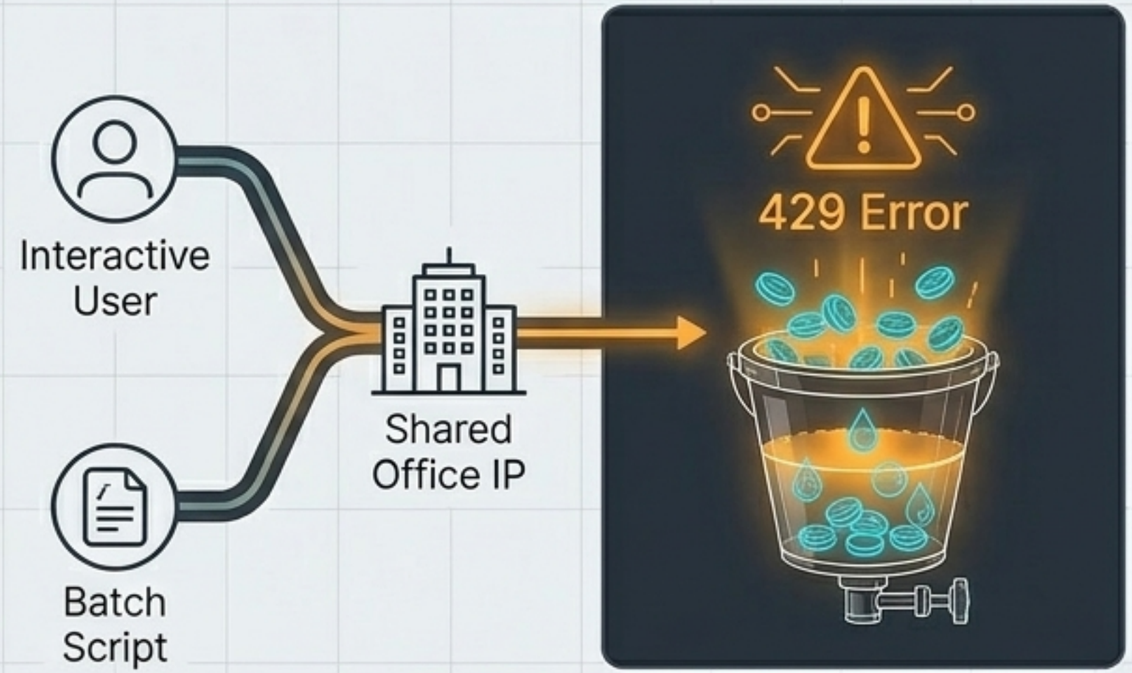
Rate Limiting Mechanics: The Token Bucket



Protects upstream servers from floods and protects LLM API budgets from runaway autonomous agents.

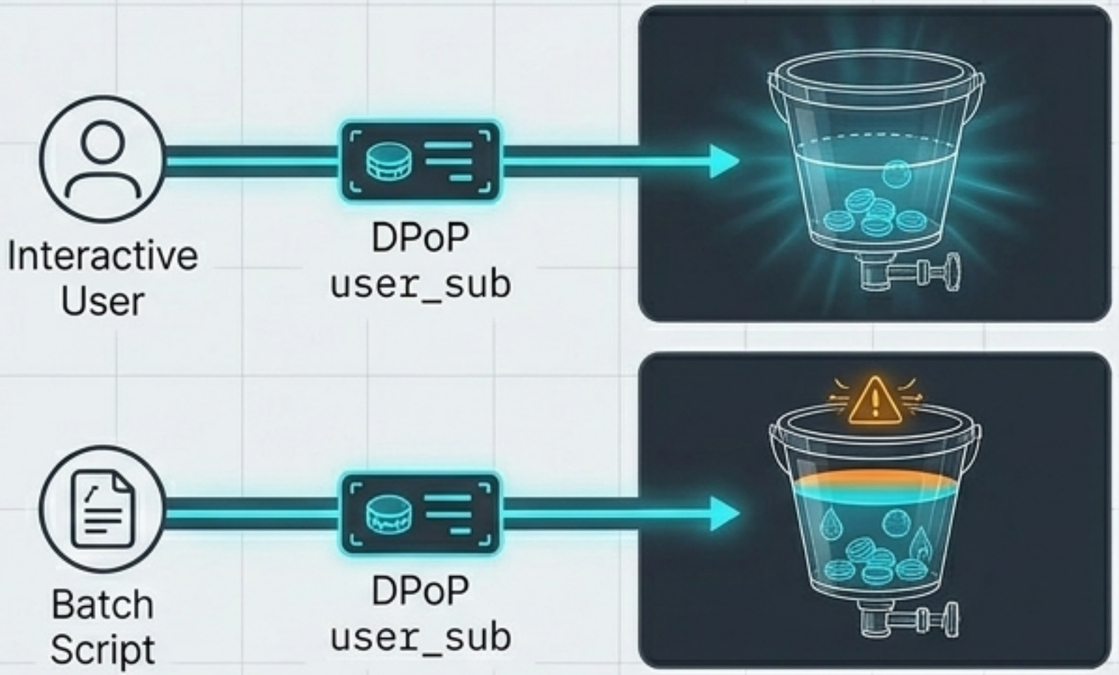
Protecting the LLM Budget: Identity vs. IP

IP-Based Routing



An interactive user crashes because a colleague's batch job emptied the shared bucket.

Identity-Based (user_sub)



Rate limits keyed on `user_sub`. Each user gets an independent budget regardless of network origin.

Differentiate burst profiles via `client_id`:

- `Interactive Users (Burst: 10)`
- `Batch CI/CD Scripts (Burst: 50-100)`

The SRE Diagnostic Manual

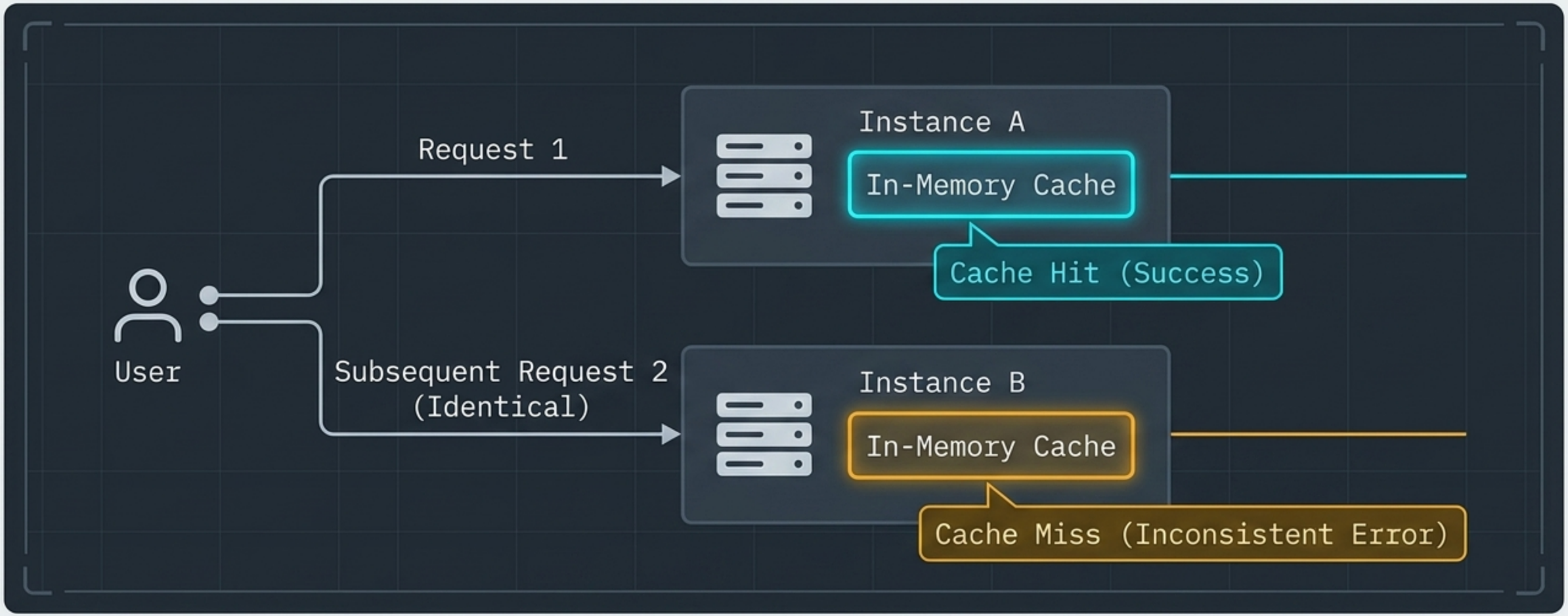
Architecture is how you build it.
Operations is how it breaks.

The 5 failure modes that **destroy production deployments**—and how to mitigate them.

Failure Modes 1 & 2: Sessions and SPOFs

Failure Mode (What happens)	Mitigation Strategy
1. Token Expiry Mid-Session DPoP token expires (15-60m). Session dies silently; user sees blank tools. 	**Proactive Refresh** Client must implement proactive token refresh when expires_in < 60s. 
2. Gateway Single Point of Failure The single gateway instance crashes at 2 AM. All traffic fails. 	**Active-Active HA Proxying** Run 2+ stateless gateway instances behind an Nginx/HAProxy load balancer with `/health` checks. 

Failure Mode 3: The State Leak

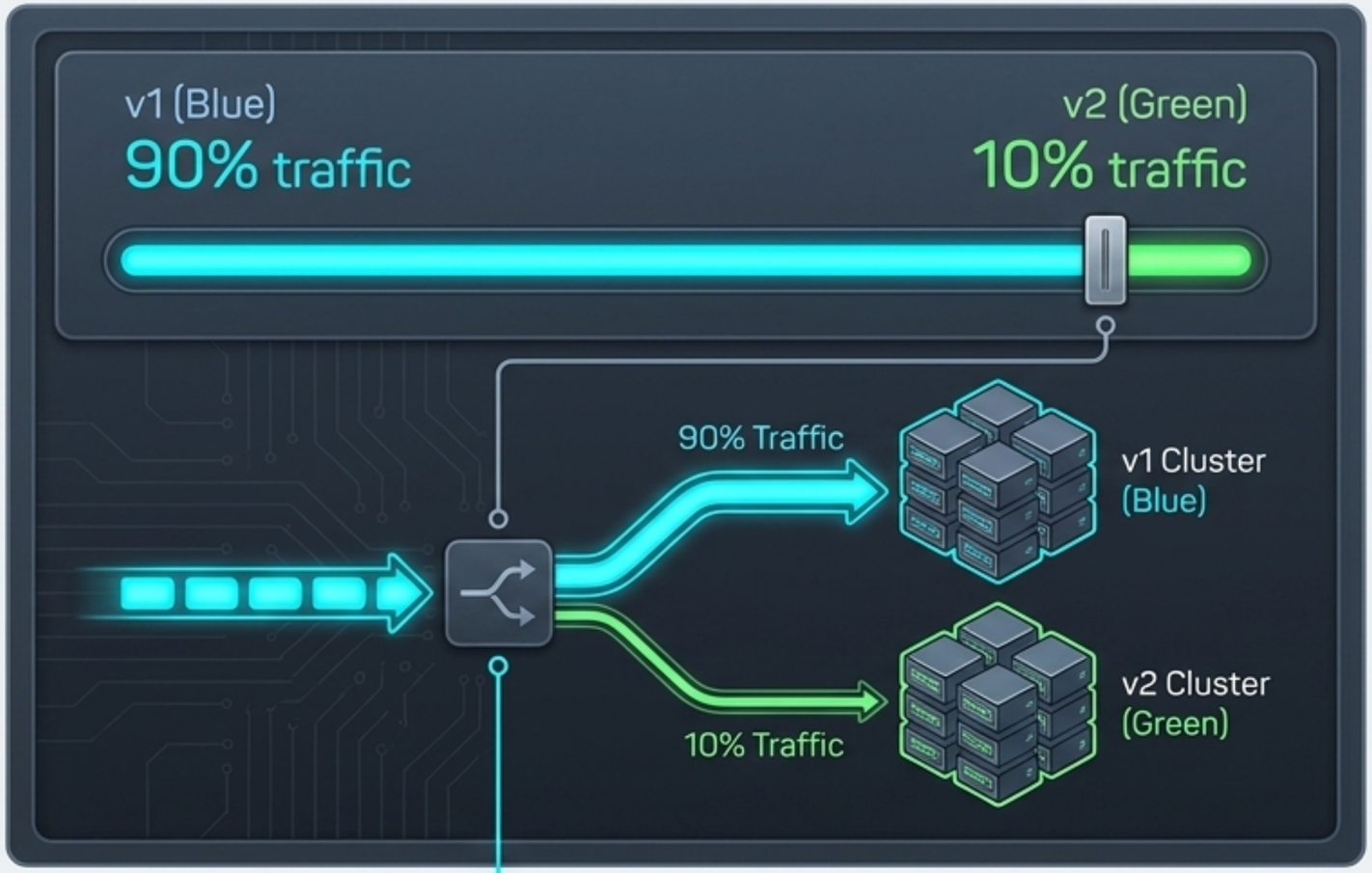


Mitigation: MCP servers must be strictly stateless across requests. Move state out-of-process to an external cache (Redis) keyed on user_sub + args hash.

Failure Modes 4 & 5: Logs and Limits

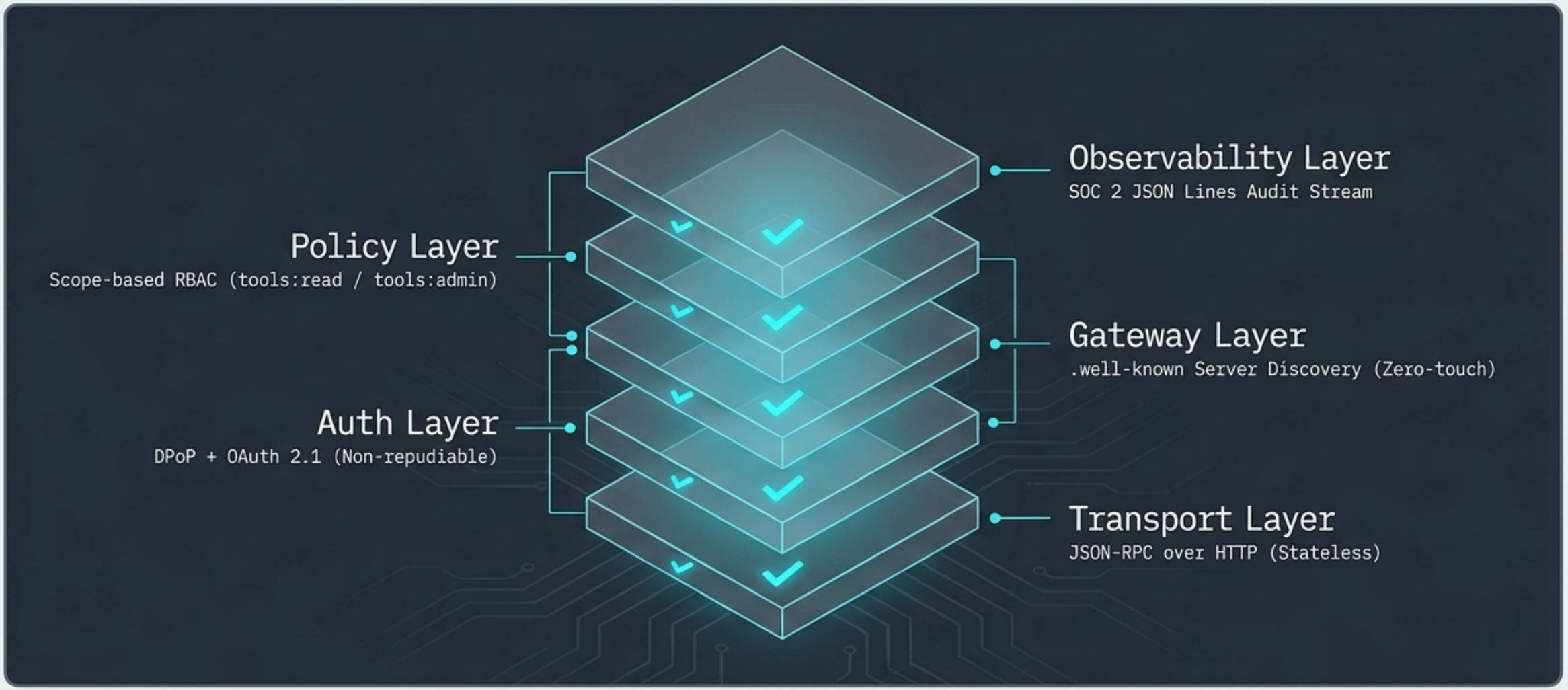
Failure Mode (What happens)	Mitigation Strategy
4. Audit Log Storage Saturation 1,000 users = 26 GB/day. Local /var/log partition fills up and crashes the server. 	**Stream to Observability Backend** Never write to local files in production. Stream JSONL directly to Loki, Datadog, or Splunk. 
5. Rate-Limit False Positives A legitimate user batch-processing files hits the burst cap and crashes mid-script. 	**Client-Specific Allowances** Design specific burst headroom allowances based on token client_id (Interactive vs. Batch). 

Horizontal Scaling & Zero-Downtime Deployments



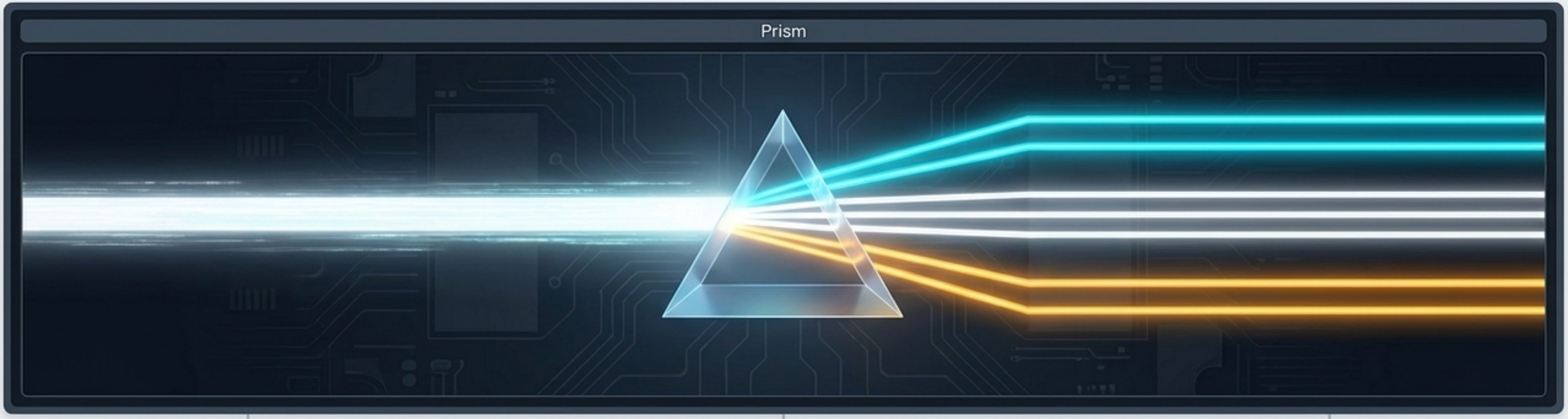
Because the JSON-RPC wire protocol and the upstream servers are entirely stateless, the gateway can seamlessly shift traffic. Start at 10%, monitor error rates in the JSONL audit stream, then ramp to 100% and drain v1.

The Production MCP Stack Built



This is the blueprint for securely operating agentic workflows at an enterprise scale.

Architectural Principles for Scale



> 1. Route at the Edge.
Put the gateway in front on Day 1.

> 2. Enforce via Scopes.
Never embed authorization logic inside individual servers.

> 3. Scale Without State.
In-memory data is the enemy of horizontal deployments.

The Model Context Protocol connects LLMs to your infrastructure.
The Gateway protects your infrastructure from the LLMs.