Google Open Source

Architecting Internet-Scale Agent Skills with Managed MCP

Prashanth Subrahmanyam

APAC lead, Developer Adoption

Google Cloud

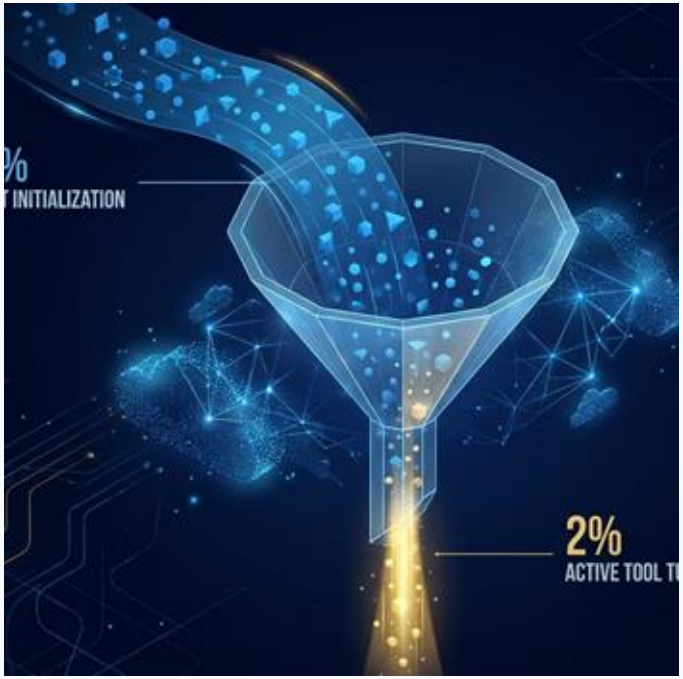
The Production Cliff & Sprawl

Rigid Boundaries: Firewalls, compliance mandates, and private VPCs shatter local connection scripts.

Drop-Off Telemetry: Only of initialization events ever convert to active tool execution.

The Vanity Fallacy: Server connection-init events do not equal functional, healthy agent workflows.

Agentic Sprawl: Linear growth in agent counts triggers exponential unmanaged database connection pools.



The Stateful Connection Bottleneck

Persistent Sockets: Stateful STDIO and SSE transports bind session state to a specific physical container node.

Chatty Traffic: A single multi-turn reasoning agent loop generates high traffic overhead:

> ~ 100 JSON-RPC messages per tool turn

The Load-Balancer Trap: Re-routing subsequent turns across scaled replicas breaks if the session context is locked inside a different VM's RAM.

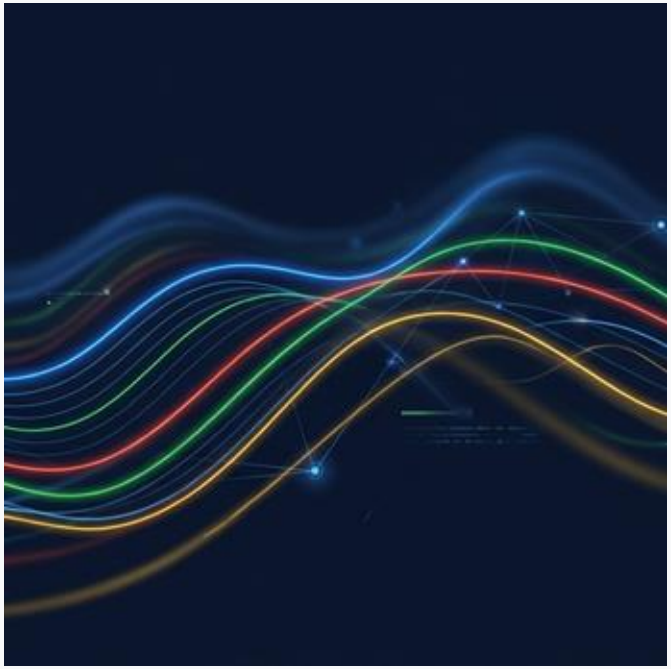
The Platform Case for Cloud Scale

Isolated Sandboxes: Run untrusted code scripts safely inside highly responsive, VPC-bound secure sandboxes.

Decoupled Memory Banks: Partition short-term session flows from long-term profile memories.

Unified Discoverability: Catalog agent capabilities inside a centralized, searchable registry.

Central Governance: Govern sprawl by establishing common auditing, compliance, and risk baselines.



Stateless Transports: SEP-1442

No handshakes: Deprecates the mandatory `initialize` handshake, folding negotiation directly into the first request.

Explicit State Handles: Memory state is returned dynamically to the client as an opaque handle.

No Affinity Lock: Requests are completely self-contained. Any server replica can parse and execute any request.

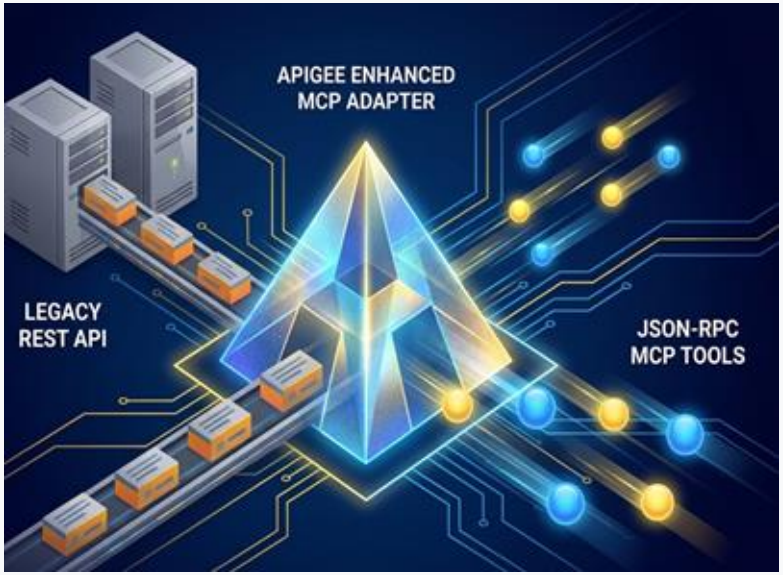
Elicitation & Header Routing

Multi-Round Trip Requests (SEP-2322): Mid-operation prompts return InputRequiredResult with an encrypted requestState token. The client responds, and any replica decodes the token to resume.

Header Elevation (SEP-2243): Routing keys are promoted directly to HTTP Headers, including:

```
- Mcp-Method: tools/call
- Mcp-Name: spanner.execute_sql
```

Zero Payload Parsing: High-performance proxies route requests instantly at the network edge without reading heavy post bodies.



No-Code Legacy Integration

No-Code REST Mapping: Apigee translates JSON-RPC payloads into existing, secure REST APIs natively.

Semantic Caching: LLM-optimized caching detects prompt intent, immediately serving cached results.

Backend Safeguards: Envoy-level rate limiters and quotas shield production systems from chatty agent loops.

Enterprise Security: Inherit corporate OAuth, private VPC compliance, and DLP patterns out-of-the-box.

Managed and Remote MCP

Hosted Endpoints: Native serverless hosting of remote MCP servers for AlloyDB, BigQuery, and Google Maps.

SPIFFE Identity: Automated assignment of cryptographically-attested SPIFFE IDs to every agent.

Unified Access: Call Remote MCPs easily from standard tools using the open-source ADK 2.0 framework.

Let's Build the Future

"Let's move past local, connection-bound scripts. Let's build a global, interconnected network of enterprise-grade, stateless skills."

Developer Workspace

antigravity.google

Open-Source ADK 2.0

adk.dev

Google Open Source

Thank you.

Prashanth Subrahmanyam

LinkedIn: in/ksprashu

X: @ksprashu

Medium: @ksprashu