
MCP DEV SUMMIT
OPERATING MCP IN THE ENTERPRISE

AUTHORS

NEELABH TRIPATHI

Engineering Architect at Cisco

AMAR DEEP SINGH

AVP IT Architecture at GM Financial

KEY CONCEPTS

- ✓ LLMs generate text, reason over context, and produce structured tool-call decisions.
- ✓ Vectors help AI systems to transform, chunk, store, and retrieve data using semantic embeddings for intelligent search.
- ✓ Tools are external functions and APIs that enable models to take real-world actions.
- ✓ Agents are AI systems that plan, decide, and act using models and tools.

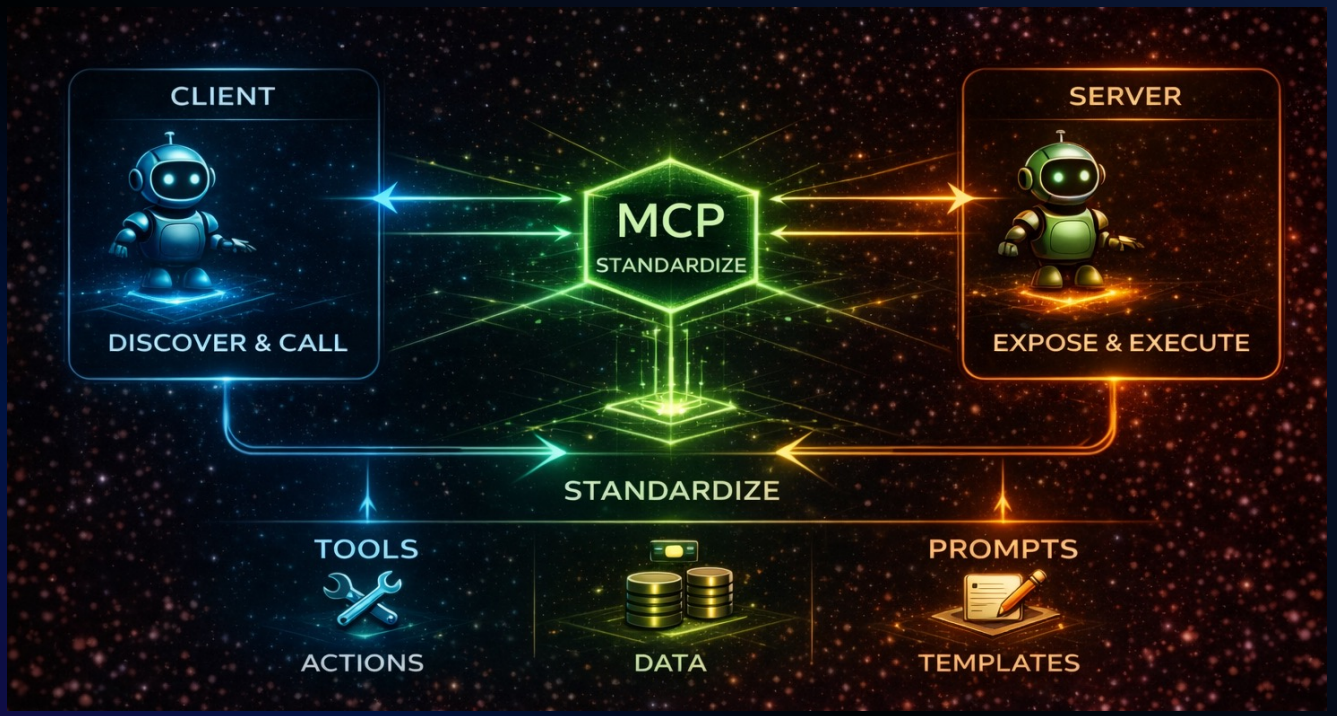
WHAT IS MCP

MCP is a standardized protocol that enables AI applications to securely connect, discover, and interact with external tools, data, and systems.

- ✓ MCP is intent-driven and dynamically discovers capabilities.

WHAT IS MCP

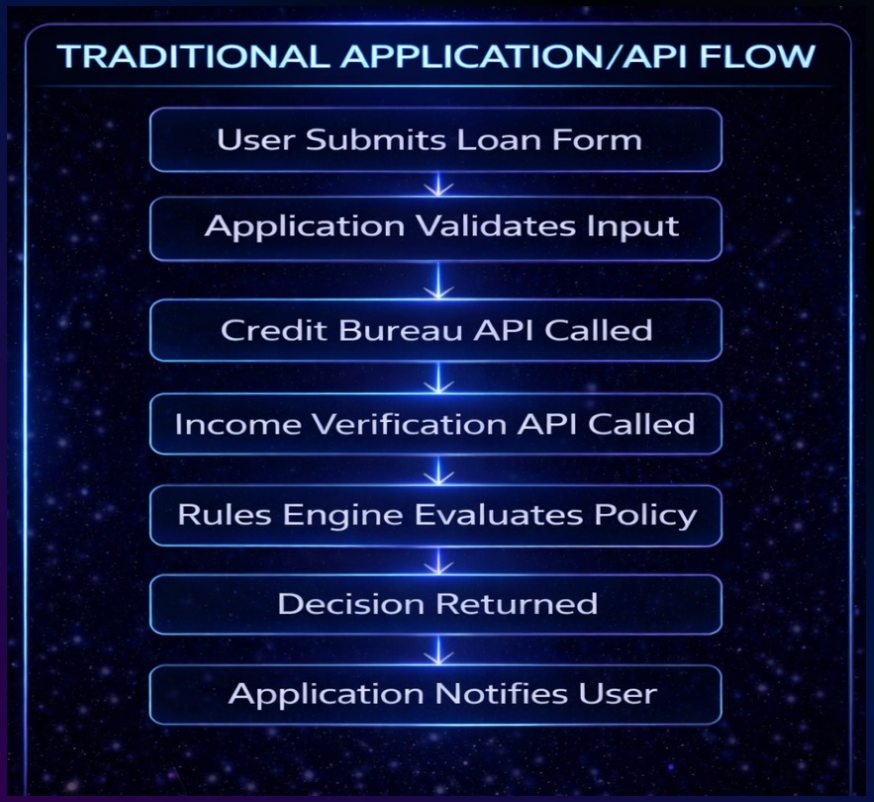
- ✓ Client discovers the tools
- ✓ Server exposes the tool
- ✓ MCP protocol standardize the communication



KEY CONCEPTS

WHAT IS MCP

Traditional Application flows are deterministic while agents are intent driven. Traditional application flows are designed in advanced but for agent goal is fixed but flow is dynamic.



V
S

FROM EXPERIMENTATION TO PRODUCTION

MCP discussions focus on protocol design, not enterprise operations.

Fragmented AI tools

AI capabilities are rapidly growing across teams, but without a platform, tools are adopted in silos—leading to duplication, inconsistent patterns, and lack of standardization.

Lack of observability

AI-driven workflows are dynamic and non-deterministic, making it difficult to trace decisions, monitor behavior, and debug failures.

Integration with legacy platforms

Enterprises operate complex ecosystems of legacy systems, APIs, and data sources that are not designed for direct AI interaction.

Security & governance concerns

AI systems interact with sensitive enterprise data and APIs, yet lack consistent enforcement of identity, policy, and compliance controls across environments.

Uncontrolled agent behavior

Agents dynamically choose tools and execution paths, which can lead to unpredictable actions, excessive access, or unintended outcomes.

Innovation in Industry

AI enables new capabilities and faster innovation, but without structure, scaling innovation leads to chaos instead of value.

PLATFORM FOR OPERATIONAL EFFICIENCY

WHAT GOES WRONG WITHOUT A PLATFORM?

Reinventing the Wheel

- Every team builds its own auth, logging, retry logic, and error handling for MCP servers
- No shared MCP server catalog — 5 teams build 5 Jira connectors with 5 different quality bars
- Cross-cutting concerns (TLS, secrets, health checks, circuit breakers) solved N times instead of once

Unobservable Systems

- Standard APM misses GenAI signals: token usage, tool latency, safety blocks
- Distributed traces break at agent-to-MCP boundary, hiding root causes
- Teams can't distinguish model failures from tool failures during incident triage

Security Blind Spots

- No tool allowlists — agents invoke any exposed tool without scope restrictions
- Prompt injection (OWASP LLM01) and unsafe output handling (LLM05) go undetected
- No forensic link between tool invocations and authenticated user principals

API Key Sprawl

- Every team creates their own tokens — no central identity, no audit trail, no rotation policies
- One compromised key exposes the entire tool chain — no way to trace the blast radius
- Compliance audits fail without a single source of truth for key-to-principal mapping

No/Unreliable Operational Resilience

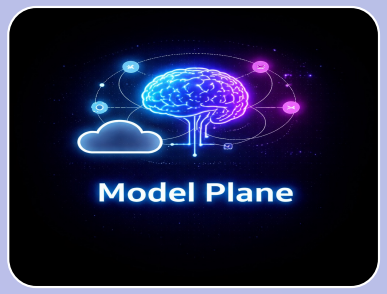
- No auto-scaling, no
- No rollback strategy Patching, certificate rotation, and dependency updates are manual, per-team, and often forgotten

Cost Blowups

- No token budgets or quotas — one runaway agent burns your monthly budget overnight
- Unmetered model calls make cost attribution and chargeback across teams impossible
- No per-agent or per-tenant spending caps to enforce financial guardrails

Without a platform, every team pays the full operational tax independently — N teams × M concerns = unsustainable complexity

PLATFORM ARCHITECTURE



Security, Governance & Observability

Identity & Access Policy, OpenTelemetry, Policy-as-Code, mTLS

AI Model Gateway

Provider routing & fallback, per-tenant budgets, token caps, audit logs

Runtime & Orchestration

Bounded workflows, tool allow lists, human-in-the-loop checkpoints

MCP Server

Domain-scoped tool microservices: Repo/Docs, ITSM, CI/CD, Knowledge/Data

SECURITY & GOVERNANCE

TOOL MISUSE

- Tool allowlists per agent — agents can only invoke servers explicitly granted to them
- Human-in-the-loop gates for destructive operations (deploy, delete, IAM changes)
- Scoped permissions: read-only vs read-write per tool per environment

DATA LEAKAGE

- Context filtering at the MCP server layer — strip PII before it reaches the model
- Output validation on model responses — block sensitive data from surfacing to users
- Data classification labels enforced per resource — Confidential data stays in Confidential tools

PROMPT INJECTION

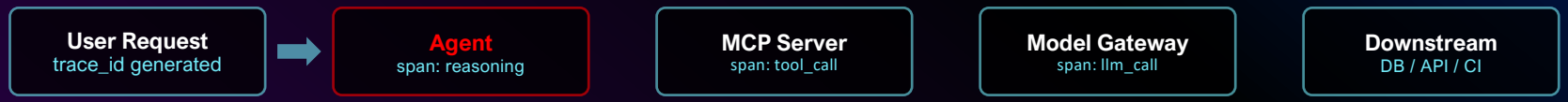
- Input sanitization at MCP server boundaries — validate and escape before tool execution
- Separate system prompts from user context — treat tool outputs as untrusted data
- OWASP LLM01/LLM05 mitigations built into the platform, not left to individual teams

IDENTITY & AUDIT

- Every tool call tied to an authenticated principal — user → agent → tool chain fully traced
- mTLS between all planes — no plaintext service-to-service communication
- Immutable audit logs for compliance — SOC 2, HIPAA, FedRAMP-ready from day one

Policy-as-Code · Identity-Aware MCP Servers · Tool Allowlists · Context Filtering · Immutable Audit Trails

OBSERVABILITY FOR AGENT SYSTEMS



DISTRIBUTED TRACING

- OpenTelemetry spans across every plane boundary
- Single trace_id from user request through agent, tool calls, and model inference
- Distinguishes model failures from tool failures during incident triage

GENAI-NATIVE METRICS

- Token usage per agent, per tenant, per model — real-time cost attribution
- Tool invocation latency, success/failure rates, and safety-block counters
- Model gateway p50/p99 latency, fallback trigger rate, throttle events
- Standard APM may not capture these

AUDIT & ALERTING

- Every prompt → completion pair logged with principal, timestamp, and tool context
- Anomaly detection: runaway loops, budget spikes, unusual tool call patterns
- SRE dashboards: agent health, tool availability, model provider status

Without observability, agent systems are black boxes — you can't debug what you can't see

LAPTOP TO PLATFORM: THE ENTERPRISE GAP

OAUTH 2.1 & PKCE

- MCP spec mandates PKCE for all auth flows — designed for browser-based clients, not headless backend services
- Client Credentials grant was removed then re-added (2025-11-25 spec) — still a draft extension
- Event-driven agents need pre-consented delegation tokens or workload identity + impersonation

TRANSPORT: STDIO → STREAMABLE HTTP

- Local stdio is near zero-latency. Remote MCP servers need DNS, TLS, load balancing — even from laptop IDEs
- Mcp-Session-Id ties session state to specific server instances in most implementations — no standard session resumption or migration mechanism exists yet

STATE & SESSION MANAGEMENT

- Any MCP server behind a load balancer hits this - laptop client or backend agent, the server side is the same
- MCP sessions are tied to transport connections - session dies when connection drops
- Sticky sessions or external session stores needed - spec has no standard serialization for session state

IDENTITY PROPAGATION

- On a laptop, agent IS the user. On a platform, agent acts ON BEHALF of a user via service identities
- Token exchange layer needed: user PKCE token → scoped delegation token → agent runtime
- MCP spec defines server as both resource + auth server — hard to integrate existing IdPs

MULTI-TENANCY & SCALING

- Laptops are single-tenant. Platform needs tenant isolation, resource quotas, noisy-neighbor protection
- MCP server cold-starts/autoscaling break interactive feel when scaled down.
- Load balancers must parse JSON-RPC payloads to route — no standard HTTP path-based routing

The MCP spec is evolving fast — Client Credentials, session decoupling, and transport improvements are on the 2026 roadmap

QUESTIONS

Connect With Amar

Buy Me