CISCO Connect

Assure Digital Resilience Across Every Network

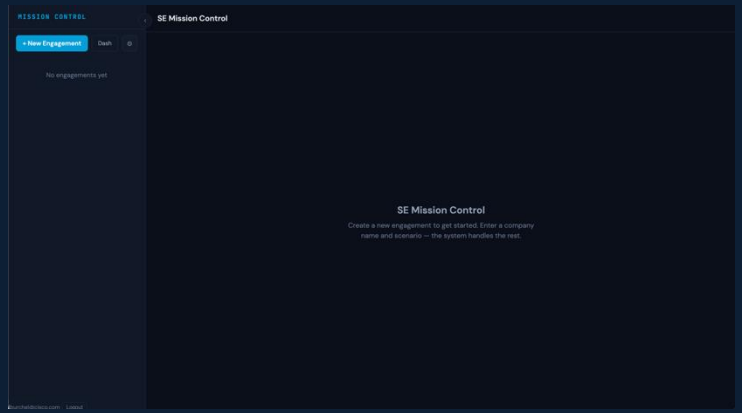
Forest Burchell- Leader, Solutions Engineering, US
Commercial & Latin America

April 7, 2026

I built an AI Application.

Mission Control

se-missioncontrol.com



An AI-powered platform built for our ThousandEyes SE teams. Used daily. In production. Real decisions. Real network calls.

- Automated Research
- Transcript Analysis
- Synthetic testing via TE API
- Automated Runbooks / Playbooks
- Work product deliverables

Then it broke.

Tuesday 2:14 PM

SILENT

AI gateway silent model downgrade

Anthropic API returns 529 (overloaded). LiteLLM triggers OpenAI failover automatically. Battle cards generated by gpt-4o-mini instead of Sonnet 4.5 -- shallower analysis, generic positioning, missed competitive nuances. SE walks into meeting with a C-grade card. No errors in logs, just a model field nobody checks.

Thursday 9:47 AM

SILENT

Guardrail Bypassed

Regional DNS for guardrail intermittently times out. AI Gateways pre-request guardrail hook fails open after 3s timeout. Prompts containing private data or adversarial content go directly to Anthropic unscreened. No alert fires because the request itself succeeds. Guardrail bypass rate climbs silently until someone audits the logs manually.

The following Monday

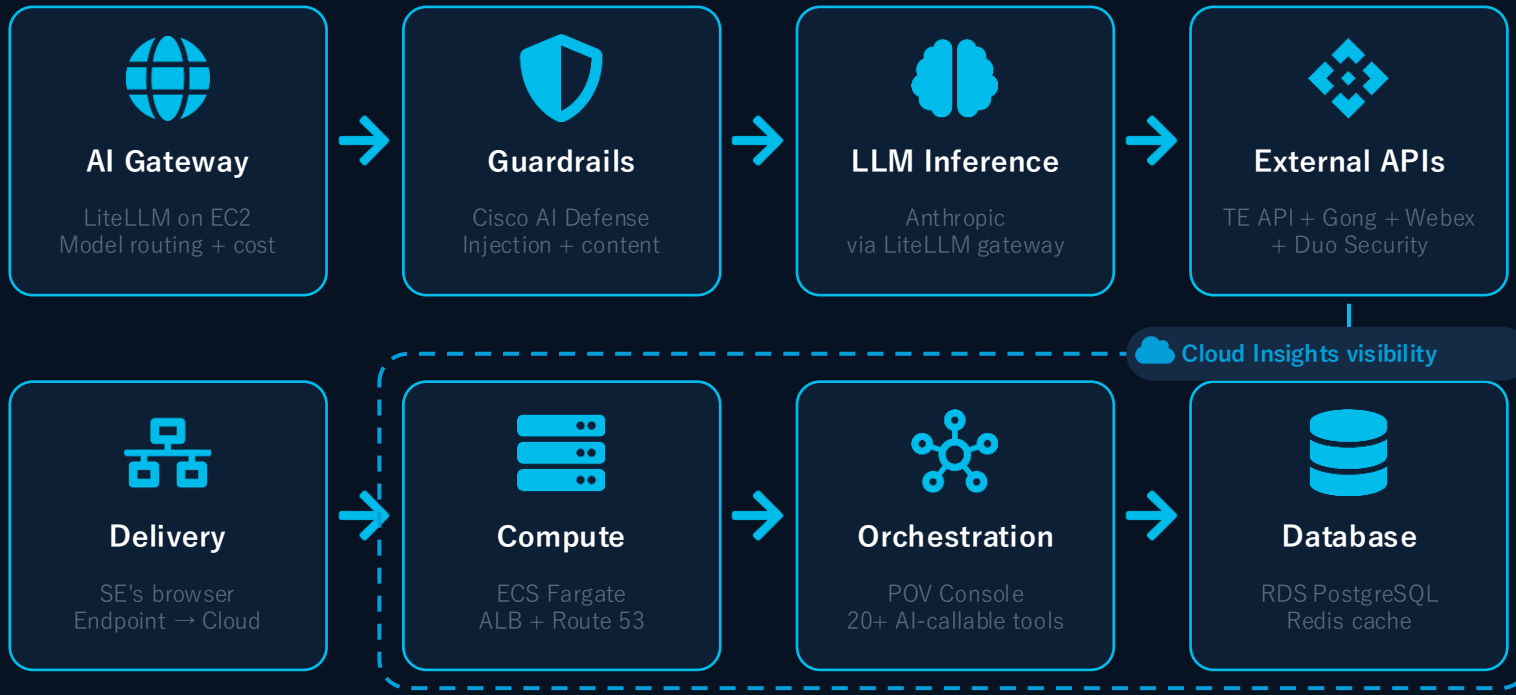
DISCOVERED

MCP tool unreachable — agent bypassed and kept going

ThousandEyes API server returned 503 during a POV setup. Agent skipped test creation step and told the SE it was complete. SE showed up to a customer meeting with zero tests deployed.

Under the Hood

One SE request to Mission Control. Seven network-dependent hops.



Every node is a network service. Every arrow is a network call. Every call can fail.

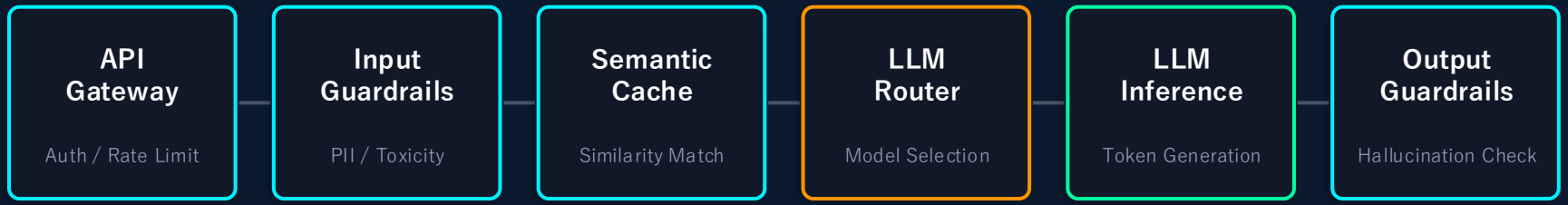
Mission Control has 12 API endpoints.

How many do you have?

Do you know?

The Invisible Middle

What actually happens between a user prompt and an AI response



This entire layer is unmonitored in most organizations

Traditional APM tracks infrastructure. LLM observability tracks tokens. Nobody tracks the orchestration that connects them.

6+	< 2%	48s
Pipeline stages per request	Orgs monitoring this layer	A single step can silently consume

Three Gaps Nobody Talks About

Where AI monitoring falls short today

Availability Without Quality

Your pipeline returns 200 OK, but the answer is fabricated. Uptime metrics say everything is fine. Your customer says otherwise.

Latency Without Context

A 4-second response looks normal. But if one pipeline stage consumed 3.8 of those seconds, you have a ticking time bomb. No tool surfaces this.

Observability Without Assurance

Token-level telemetry tells you what the model did. It doesn't tell you whether the answer was correct, grounded, or safe.

The Visibility Gap

AI applications span your infrastructure and external providers. Current tools only see half the picture.

WHAT YOU CONTROL

Your Infrastructure

- Internal networks & routing
- Cloud environments (AWS, Azure, GCP)
- Application code & agents
- API gateways & load balancers

BLIND SPOT
Internet & CDN paths

WHAT YOU DEPEND ON

AI Provider Stack

- LLM APIs (OpenAI, Anthropic, Bedrock)
- MCP Tools (ServiceNow, Jira, Webex)
- Provider CDNs & edge infrastructure
- Rate limits & capacity constraints

APM Tools — PARTIAL

See application traces but blind to network path issues

Visibility:

App traces | Code spans | Network path | Internet hops

LLM Observability — PARTIAL

Track tokens and prompts but assume the network works

Visibility

Tokens | Latency | Root cause | Path issues

Provider Dashboards — PARTIAL

Show usage stats but no diagnostic capability

Visibility

Usage stats | Status page | Your path | Diagnostics

When your AI agent fails or slows down, where do you look?

API gateway? Provider infrastructure? Internet routing? CDN issues? Regional capacity?

The AI Monitoring Maturity Model

Most organizations are stuck at Layer 1

01

Infrastructure Monitoring

Network, compute, cloud availability. You know if the server is up. You don't know if the AI is working.

Where most orgs stop

02

LLM Observability

Token usage, model latency, prompt tracing. You know what the model did. You don't know if it did it well.

Emerging

03

AI Assurance

End-to-end pipeline validation, output quality testing, grounding verification. You know the AI is correct.

The destination

What Is AI Assurance?

Moving from 'is it up?' to 'is it right?'

AI Assurance is the practice of continuously validating that AI systems produce correct, grounded, and policy-compliant outputs — not just available ones.

Output Validation

Does the response match known-good baselines?

Grounding Verification

Is the answer traceable to source data?

Pipeline Health

Is every orchestration stage performing within bounds?

Quality Metrics

Availability, correctness, and latency — all three.

ThousandEyes and AI

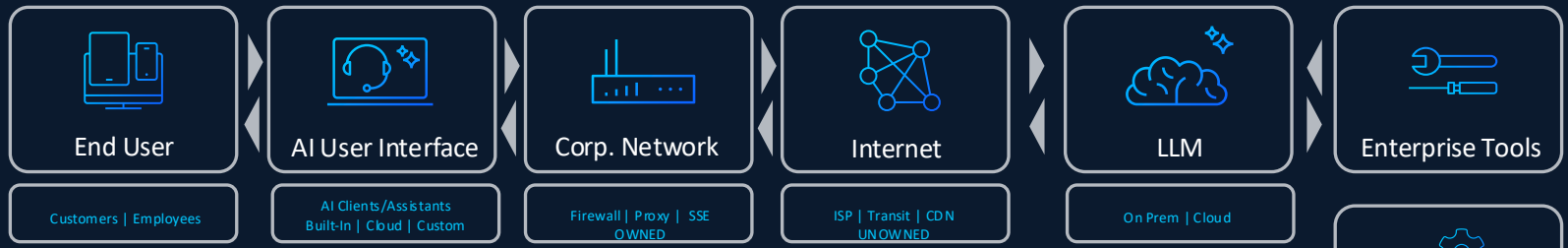
Monitor and Feed Your AI Ecosystem

- Visibility into latency, response times, token efficiency, and service availability of LLMs
- Inspect MCP resources validating the state of available tools and their configurations

ThousandEyes

Assurance for the AI Era

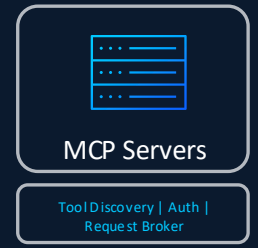
- Validate that responses maintain accuracy and consistency
- Assure connectivity across multiple owned and unowned environments



Agentic AI Ops

ThousandEyes

- ThousandEyes MCP Server with any MCP compliant client/assistant
- Conversational, AI-driven access to ThousandEyes data without context-switching.
- Faster MTTI and MTTR by replacing manual navigation/correlation with natural language queries.
- Foundational data source in modern AIOps environments – correlated with other data sources/platforms



End-to-End Test Architecture

1

API/Transaction Test

Application Logic

- Rate limit threshold alerts (80/90%)
- API response validation
- Login -> Token -> Prompt flow
- Vector Database Monitoring

2

HTTP Server Test

Connection Layer

- DNS resolution time
- TCP connect latency
- SSL/TLS handshake
- Time to first byte (TTFB)

3

Network Test

Path Analysis

- Hop-by-hop visualization
- Packet loss per segment
- Latency breakdown
- ISP/provider identification

4

BGP Monitoring

Routing Layer

- Route hijack detection
- Path change alerts
- Upstream provider issues
- Traffic blackhole prevention

So I instrumented everything.

ThousandEyes tests running against Mission Control — right now

Foundation

- DNS: se-missioncontrol.com
- DNS: LiteLLM + LLM Providers
- HTTP: ALB + LiteLLM health
- BGP Reachability

Cloud Infra

- VPC flow logs: ECS ↔ RDS, ECS ↔ Redis traffic
- Config change correlation w/ synthetic test data
- Cloud topology: ALB → ECS → security groups → RDS
- Security group + route table change detection

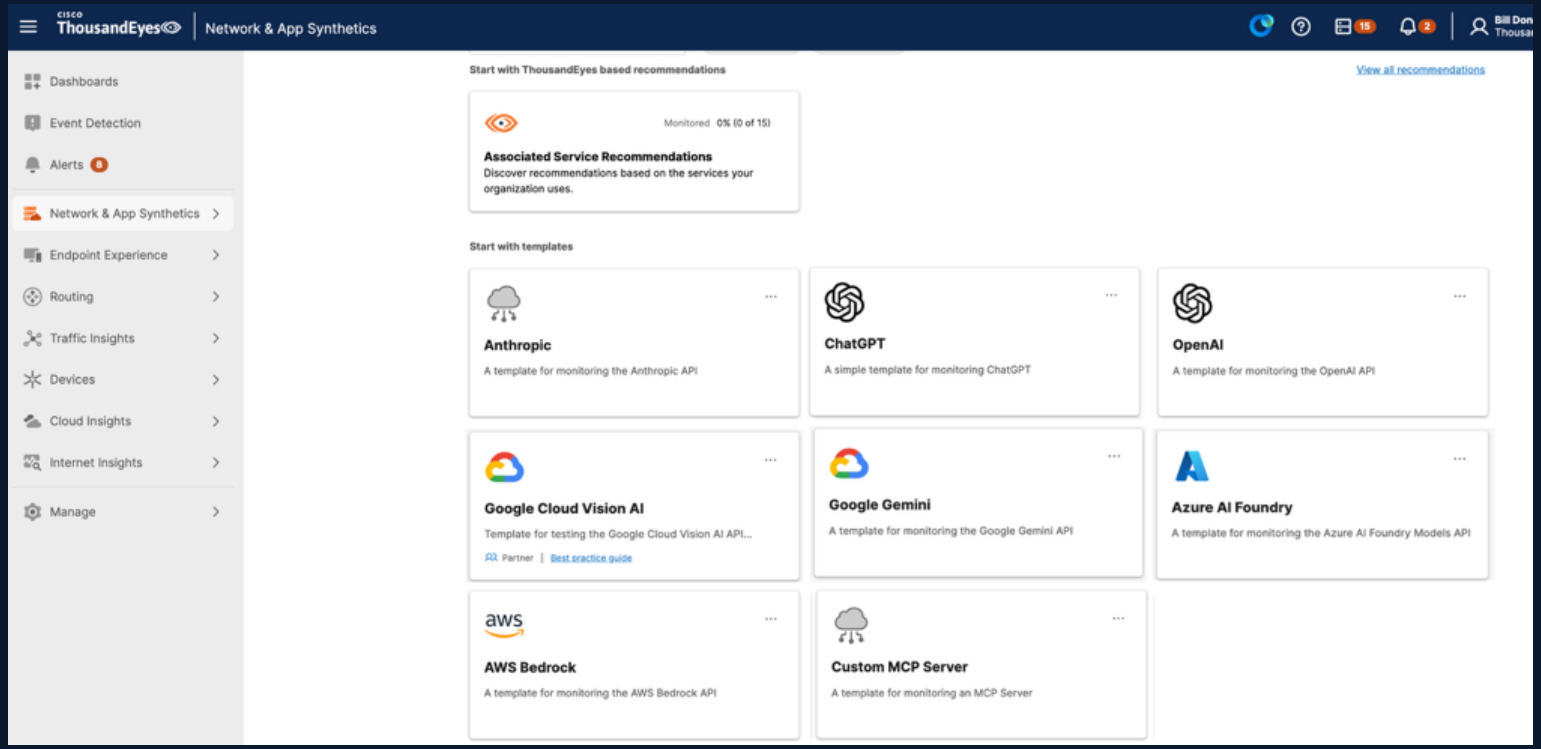
AI-Specific

- Multi-step: prompt → LiteLLM → OpenAI → validate
- AI Test Template: baseline prompt + assertions
- API chain: Gong ingest → extract → store → analyze
- Guardrail latency: < 200ms threshold

Operational

- Internet Insights: OpenAI + Duo outage detection
- Endpoint agent: SE laptop → Mission Control path
- Alert → Webex notification
- Weekly AI dependency health summary

We Can Provide AI Assurance TODAY



ThousandEyes MCP Server

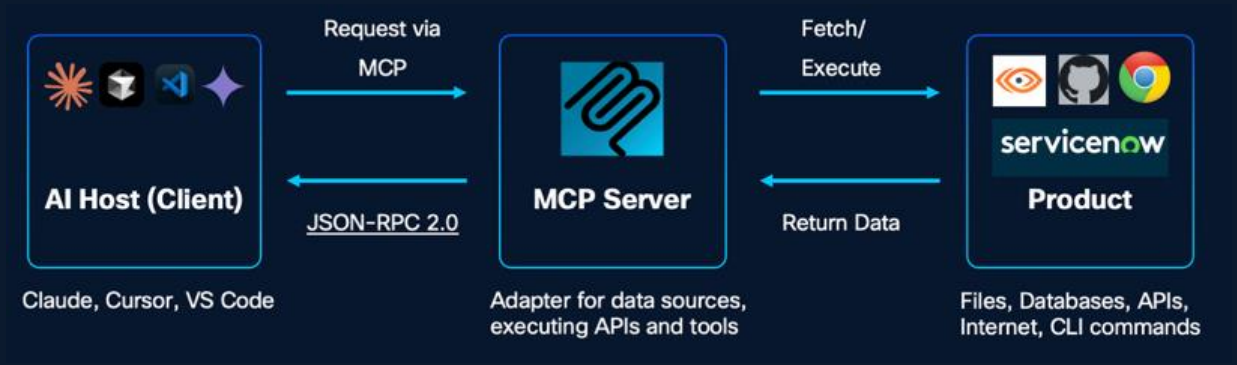
What is Model Context Procol (MCP)?

Lightweight protocol connecting AI agents to tools, APIs, databases, and filesystems

Secure, scalable client/server architecture for AI-powered workflows

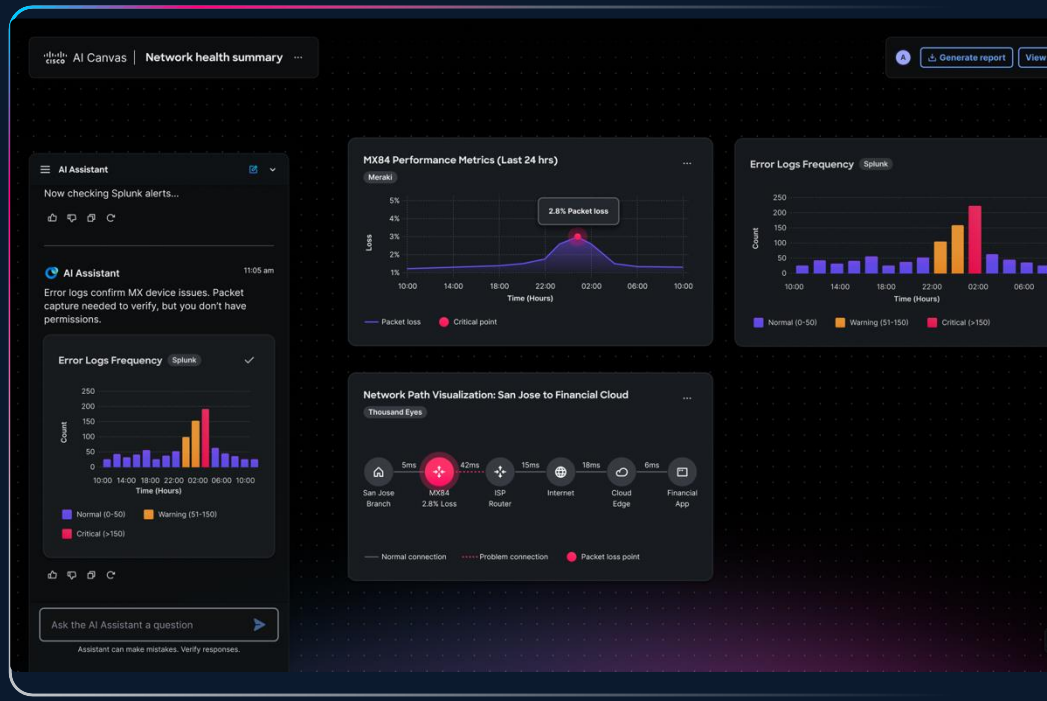
Simplifies integration by providing a uniform interface to diverse infrastructure and operational tools

Critical component for customers building Agentic AI frameworks, tools and platforms



AI Canvas

- Single canvas for cross domain troubleshooting
- Generative UI dashboard with AI reasoning built-in
- Support data from Meraki, ThousandEyes, Splunk, and more



Demo

ThousandEyes AI Assurance

Your Turn

Day 1

Monday afternoon

DNS + HTTP tests for every AI endpoint. BGP monitoring for your top 3 AI provider prefixes. Network path visualization to your primary LLM.

Cloud Insights: connect your AWS/Azure account. Topology + flow logs start immediately.

30 minutes of setup. Zero excuses.

Day 30

Validation layer

Multi-step API tests mirroring your agent workflows. AI Test Templates with baseline prompts and assertions. MCP monitoring for tool integrity.

Cloud Insights: correlate config changes with synthetic test performance. Baseline your VPC traffic patterns.

Day 90

Operational maturity

Tune alerts from 60 days of baseline data. Integrate with incident management. Cloud Insights flow log anomaly detection tuned to your environment.

This is now normal ops.

The Power of Cisco

ThousandEyes Assurance for Agentic Apps is essential as Cisco helps you build your AI Ecosystem

AI Networking

Deliver AI everywhere with intelligent networking purpose-built for high performance, simplified operations, and security at scale.

Secure AI Factory

Accelerate AI deployment with a modular and validated design that combines infrastructure, security, and observability.

AI Defense

Protect AI development and usage with visibility, guardrails and real-time threat intelligence designed for emerging AI-specific risks.

The Network Is the AI Runtime.

AI agents don't contain intelligence.
They call for it — across your network.
If you can't see every hop, you can't assure the outcome.

Track: Digital Resilience
Session Title: Assure Digital Resilience Across Every Network

We’d love to hear from you!

To tell us about your experience so far, and your thoughts on this session.

Thank you


CISCO