

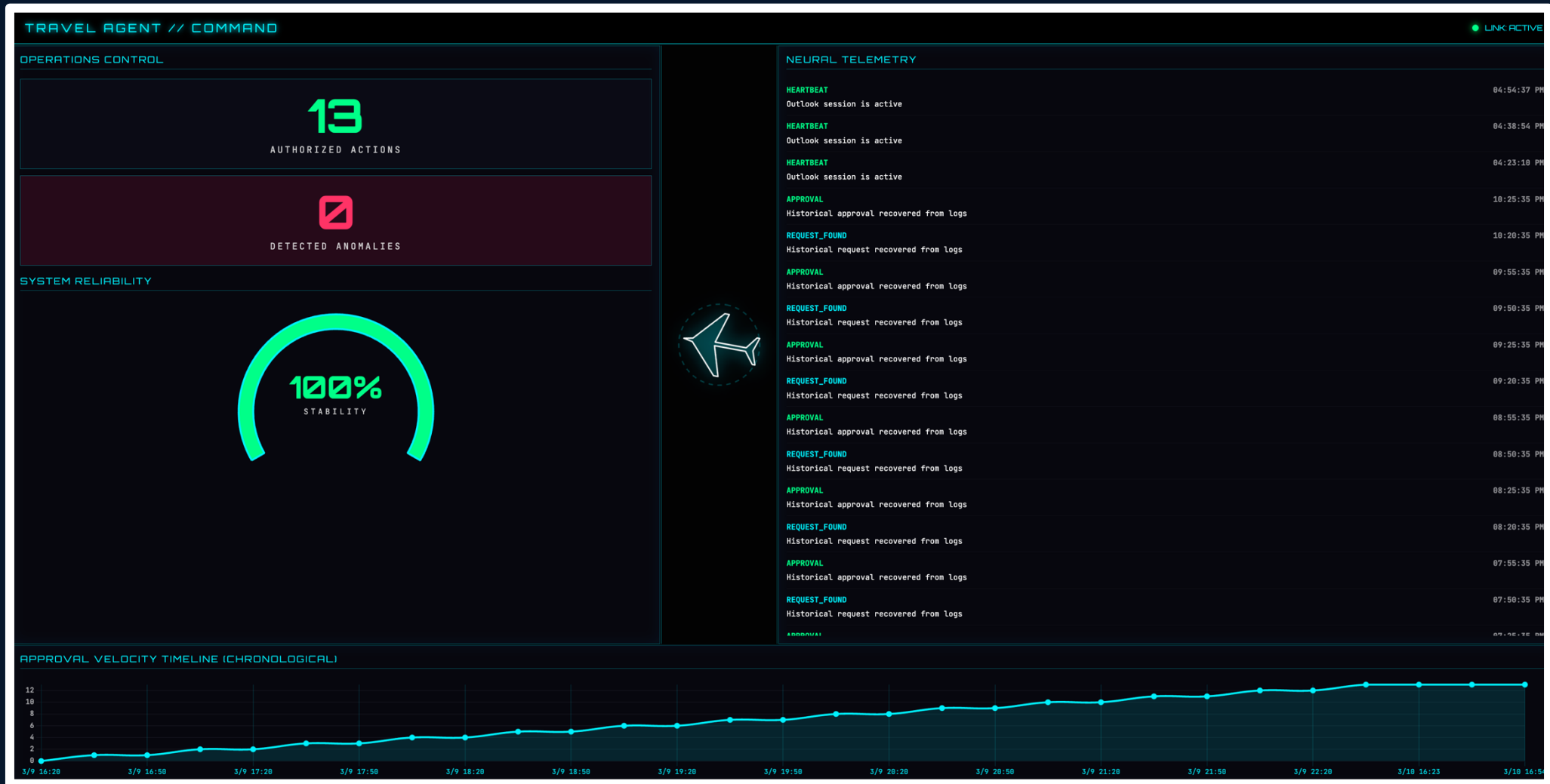
# ThousandEyes

AI Assurance and Agentic Ops

Shawn Eustis  
Director Solutions Engineering



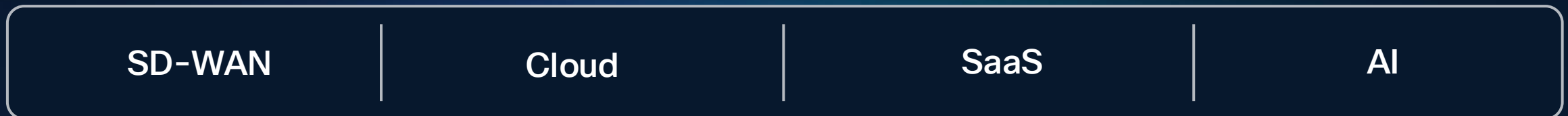
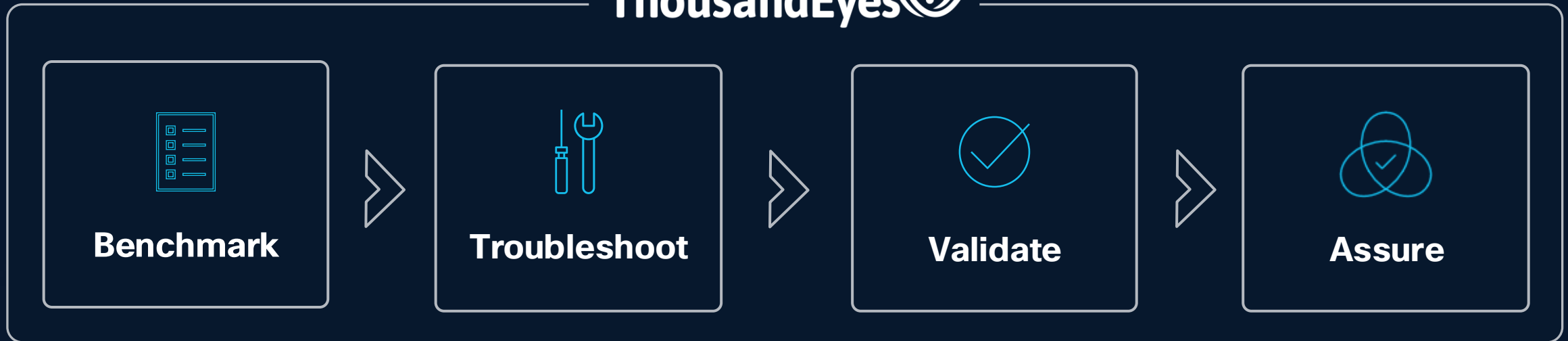
# Vacation Vibe Coding



# De-risk and Accelerate Transformative Projects

Actionable Insights Before, During, & After

ThousandEyes 

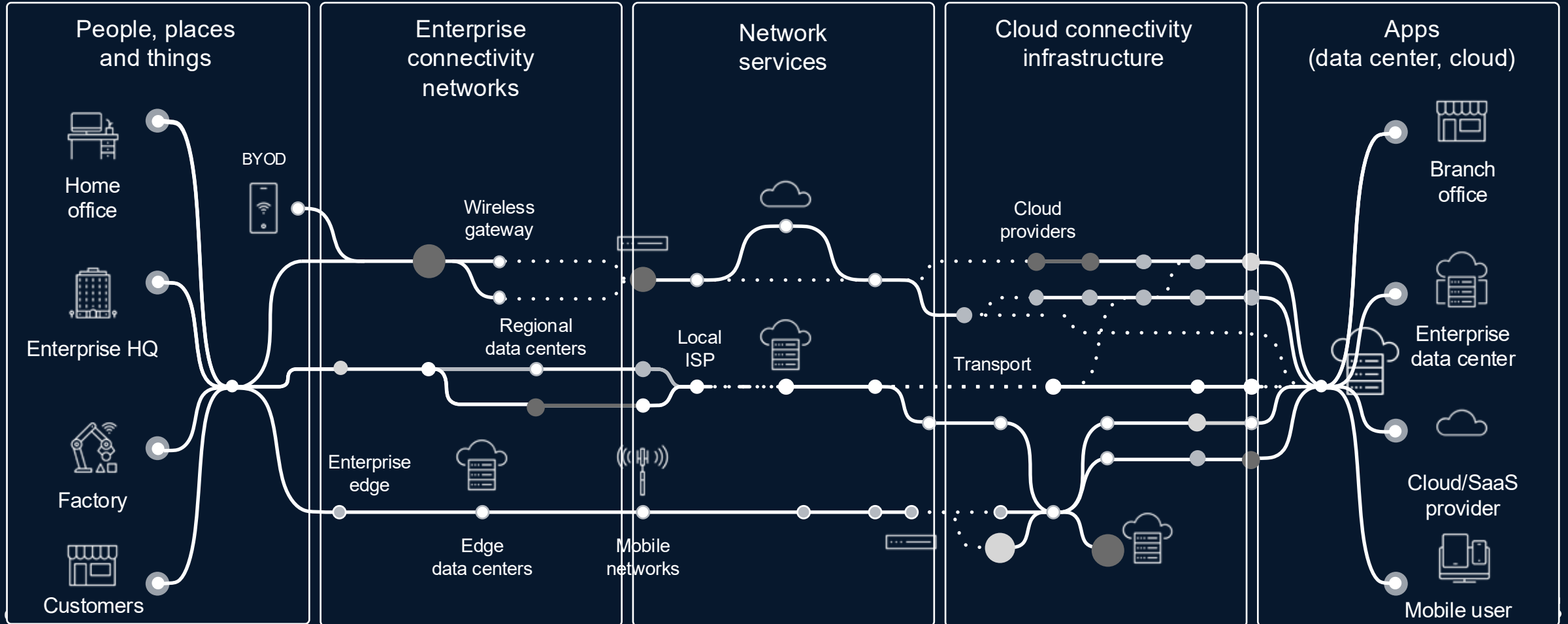


Endpoint Experience

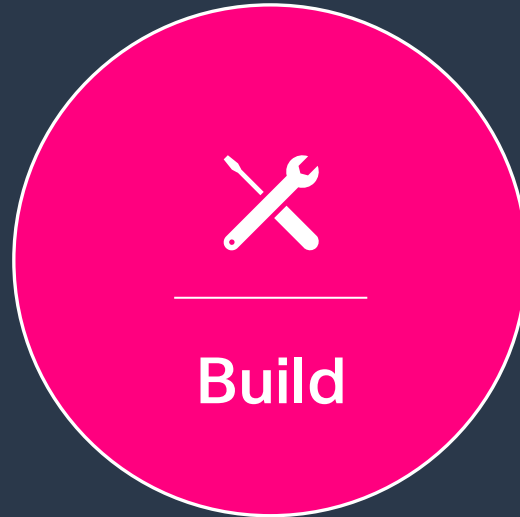
Traffic Insights

Enterprise and Cloud Synthetics

Cloud Insights



# Where are you on your AI journey?



# Why Assurance?

APPLICATION DOWN - BLAME NETWORK

NETWORK  
ENGINEER

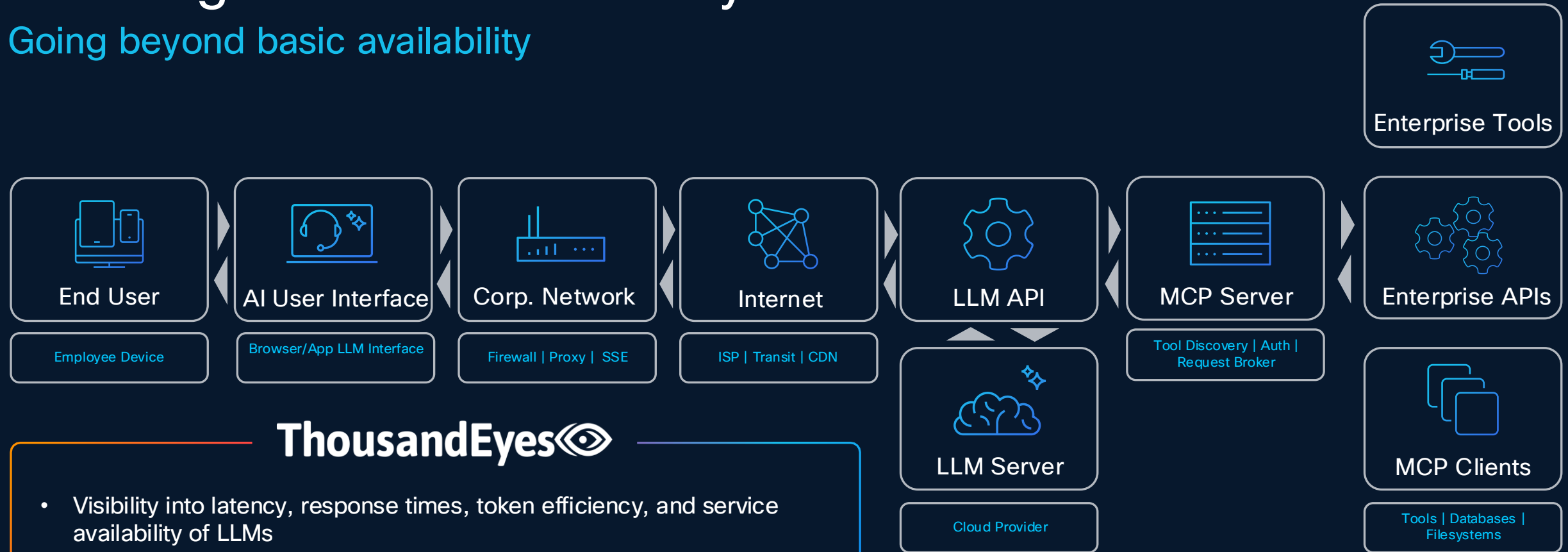






# Assuring AI with ThousandEyes

Going beyond basic availability



## ThousandEyes

- Visibility into latency, response times, token efficiency, and service availability of LLMs
- Inspect MCP resources validating the state of available tools and their configurations
- Validate that responses maintain accuracy and consistency
- Comparing performance of multiple models
- Assure connectivity across all owned and unowned environments

# The AI Monitoring Maturity Model

*Most organizations are still 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 are at

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

# 72 hours.

That's how long a production AI gave wrong answers before anyone noticed. Every dashboard said green.

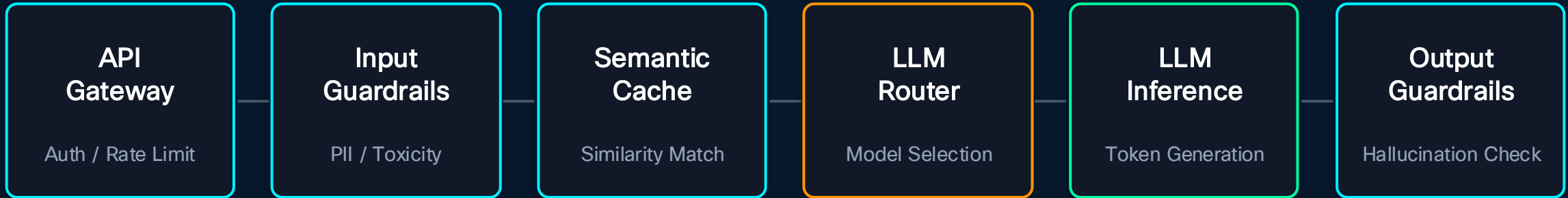
The model didn't go down. The infrastructure didn't fail.

The pipeline silently started returning confident, articulate, completely wrong answers.

*Nobody was watching the middle.*

# 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+**

Pipeline stages per request

**< 2%**

Orgs monitoring this layer

**60%**

Of query latency occurs outside the LLM inference later

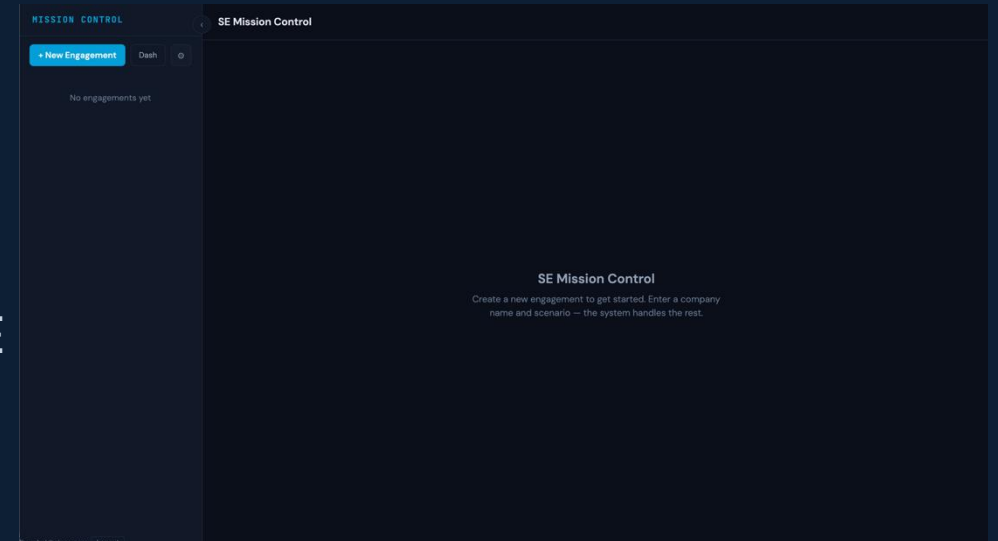
# We built an AI Application.



## Mission Control

[se-missioncontrol.com](https://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 services.



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

# So we 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



---

# Demo

## ThousandEyes AI Assurance

# Agentic AIOps - MCP Server

## ThousandEyes MCP Server



### THOUSANDEYES | Network Intelligence Platform

## Incident Report

Root Cause Analysis

Alert ID	99f6959a-b91e-4a41-ac42-0bc1413b1b1a
Test Name	Epic - Userweb Target/s: https://userweb.epic.com
Alert Started	2026-02-09 04:10 UTC
Alert Duration	15 minutes
Report Date	February 08, 2026 at 10:42 PM CST
Account ID	562949933625984
Report Type	Root Cause Analysis
ThousandEyes Link	<a href="https://www.thousandeyes.com/alerts/99f6959a-b91e-4a41-ac42-0bc1413b1b1a">https://www.thousandeyes.com/alerts/99f6959a-b91e-4a41-ac42-0bc1413b1b1a</a>

### 1. Executive Summary

A minor severity adaptive page load alert triggered at 2026-02-09T04:10:00Z (10:10 PM CST) impacting the Epic Userweb test (https://userweb.epic.com) across 9 geographically distributed cloud agents. Users experienced page load timeouts with the metric pageLoadTimeOut=true and pageLoadTimeOut=false preventing successful page completion. The incident self-resolved at 2026-02-09T04:25:00Z (10:25 PM CST) with a total duration of 15 minutes. Root cause: Simultaneous page load timeout failures at multiple US locations due to server-side responsiveness degradation at the Epic application layer. No underlying network path issues detected. This alert represents a recurring pattern occurring at approximately the same time.

### 2. Incident Timeline

- 2026-02-09T03:50:00Z (9:50 PM CST): Baseline period - 0 violations out of 11 agents (issue probability: 0.25%)
- 2026-02-09T03:55:00Z (9:55 PM CST): Violation count increased - 2 violations out of 11 agents (probability: 20%)
- 2026-02-09T04:10:00Z (10:10 PM CST): Alert triggered - 5 agents violated threshold, exceeded 100% page load timeout (probability: 100%)

Done. I've created the incident:

Incident Number: INC0080991

Field	Value
Short Description	Investigate wireless connectivity issues - MGC1DC7QFYWY on Mustang network
Category	Network
Subcategory	Wireless
Priority	1 - Critical
Urgency	1 - High

The description includes all the ThousandEyes findings: the poor RSSI readings (-67 to -72 dBm), low signal quality (54-66%), elevated roaming events indicating AP hunting behavior,

## Comprehensive Wireless Performance Report

Last 24 Hours | Browser Sessions + Scheduled Tests

SSID: corp-wifi

Overview | Browser Sessions | Scheduled Tests | Access Points

MONITORED DEVICES 2 Endpoint Agents	ACCESS POINTS 11 Unique BSSIDs	AVG SIGNAL QUALITY 72% Wireless Strength	GATEWAY LATENCY 6.8 ms Network Response	TEST AVAILABILITY 57.5% HTTP Tests
APP SCORE 89.0 Performance Index	AVG THROUGHPUT 2.9 Mbps Download Speed	HTTP RESPONSE 245 ms Test Latency		

### Executive Summary

**Network Infrastructure:** 2 endpoint devices connected to SSID "corp-wifi" through 11 different access points (BSSIDs), indicating good wireless coverage and roaming capability.

**Wireless Performance:** Signal quality averaged 72% with acceptable RSSI levels (-54 to -70 dBm). Gateway latency is excellent at 6.8ms with minimal packet loss (0.02%).

**Application Performance:** Scheduled HTTP tests show 57.5% availability with average response times of 245ms. Application score improved throughout the day, peaking at 89.0.

## ThousandEyes Service Health Dashboard

Alert Severity: 8 (8 (75.0%), 1 (12.5%), 1 (12.5%), 0 (0.0%)

Test Status: 8 (8 (75.0%), 1 (12.5%), 1 (12.5%), 0 (0.0%)

Test Type: 8 (8 (75.0%), 1 (12.5%), 1 (12.5%), 0 (0.0%)

Outage Types: 12 (12 (100%), 0 (0.0%), 0 (0.0%), 0 (0.0%)

Severity	Count	Percentage
Critical	8	75.0%
Major	1	12.5%
Minor	1	12.5%
Info	0	0.0%

Test Type	Count	Percentage
DNS Server	3	37.5%
HTTP Server	3	37.5%
Page Load	2	25.0%
Unknown	1	12.5%

Outage Type	Count	Percentage
Application	8	66.7%
Network	4	33.3%

### Active Alerts

Severity	Test Name	Type	Alert Rule	Target	Started	Duration	Violations	Actions
Critical	Production - Bugfix	Page Load	APPLICATION AVAILABILITY PROBLEM	-	Jan 7, 2026, 12:30 AM	15d 9h	30	Details
Critical	Production - Bugfix	DNS Server	NETWORK PACKET LOSS DETECTED	net.production.net	Jan 7, 2026, 12:30 AM	15d 9h	28	Details

## Microsoft Office 365 Performance Dashboard

Real-time monitoring powered by ThousandEyes | Account Group: 106370

CURRENT LATENCY 47.9 ms	AVERAGE LOSS 0.31 %	AVAILABILITY 99.97 %	RESPONSE TIME 144 ms
----------------------------	------------------------	-------------------------	-------------------------

### Network Latency Timeline

Graph showing latency (ms) over time.

### Packet Loss Timeline

Graph showing packet loss (%) over time.

### Network Path Trace - Office 365

Step	Latency (ms)	Loss (%)
1	1.2ms	0.0%
2	5.4ms	0.0%
3	12.8ms	0.0%
4	18.3ms	0.0%
5	24.7ms	0.0%
6	28.1ms	0.0%
7	32.5ms	0.0%
8	35.2ms	0.0%



# AI Assurance Demos

NAME	SCENARIO	LINK TO DEMO
Anthropic – Token Exhaustion	In a multi provider/model environment having an early warning system in place to be notified when token or request exhaustion is approaching is key for capacity and failover planning.	<a href="https://ciwlpuvdopteilegxzaszwcixfxquiwz.share2.thousandeyes.com">https://ciwlpuvdopteilegxzaszwcixfxquiwz.share2.thousandeyes.com</a>
Anthropic – Internet Outage	Agentic Applications do not contain intelligence. They call for it across the network. Network visibility from source to destination is key to understanding how the underlying infrastructure impacts Agentic performance. In this example a lumen outage leads to Anthropic being unreachable and initiating a provider failover via the AI Gateway.	<a href="https://crzleyggjtjvfcbyawetlmljvikltgzmq.share2.thousandeyes.com">https://crzleyggjtjvfcbyawetlmljvikltgzmq.share2.thousandeyes.com</a>
Anthropic – Provider Infrastructure Overloaded	The Anthropic API is returning 529 status codes, which Anthropic uses to indicate infrastructure overload, essentially telling clients that "we are at capacity, please retry." This is distinct from a standard 503 (Service Unavailable) and signals deliberate rate-limiting or capacity management on Anthropic's side. This inevitably leads to additional token consumption and agentic agent failures.	<a href="https://aouvouttttkrsfwzkgpxcphpfjmhfa.share.thousandeyes.com">https://aouvouttttkrsfwzkgpxcphpfjmhfa.share.thousandeyes.com</a>
Guardrail Impacting Agent Performance	Lakera Guard keeps latency tight and predictable across every workload, with sub-40ms for short prompts but that's the scan time. The 400+ms the sharelink shows is the network path to the service, which neither Lakera nor APM tools can see. That gap is pure ThousandEyes territory. That means 800+ms round trip (input+output scans) leading to 8x SLA impact.	<a href="https://ctavzhxqdnyaiqdbjglucnnjllavcjvm.share2.thousandeyes.com">https://ctavzhxqdnyaiqdbjglucnnjllavcjvm.share2.thousandeyes.com</a>

# You Can Get Started Now

The screenshot displays the Cisco ThousandEyes Network & App Synthetics dashboard. The top navigation bar includes the Cisco ThousandEyes logo, the page title "Network & App Synthetics", and user information for "Bill Don ThousandEyes". A left sidebar lists navigation options: Dashboards, Event Detection, Alerts (8), Network & App Synthetics (selected), Endpoint Experience, Routing, Traffic Insights, Devices, Cloud Insights, Internet Insights, and Manage. The main content area is titled "Start with ThousandEyes based recommendations" and features a card for "Associated Service Recommendations" with a "Monitored 0% (0 of 15)" status. Below this is a "Start with templates" section containing eight cards for monitoring various services: Anthropic, ChatGPT, OpenAI, Google Cloud Vision AI, Google Gemini, Azure AI Foundry, AWS Bedrock, and Custom MCP Server. Each card includes a logo, the service name, and a brief description of the template.

# End-to-End Test Architecture

1

## API/Transaction Test

### Application Logic

- Custom token utilization checks
- 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

# This Is a Business Conversation

*AI failures aren't technical incidents – they're business risk events*



## Regulatory Exposure

AI regulations (EU AI Act, NIST AI RMF) require organizations to demonstrate ongoing monitoring of AI system outputs. 'We check if it's up' doesn't satisfy compliance.



## Brand & Trust Risk

A single hallucinated answer in a customer-facing AI application can erode years of trust. Unlike a website outage, AI failures feel personal – the system gave wrong information with confidence.



## Operational Cost

Without per-stage visibility, teams spend 4-6x longer diagnosing AI pipeline issues. Every minute of MTTR is multiplied across the organization.

# 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.



Monitor



Validate



Assure

