cisco Engage

ThousandEyes

AI Assurance and Agentic Ops

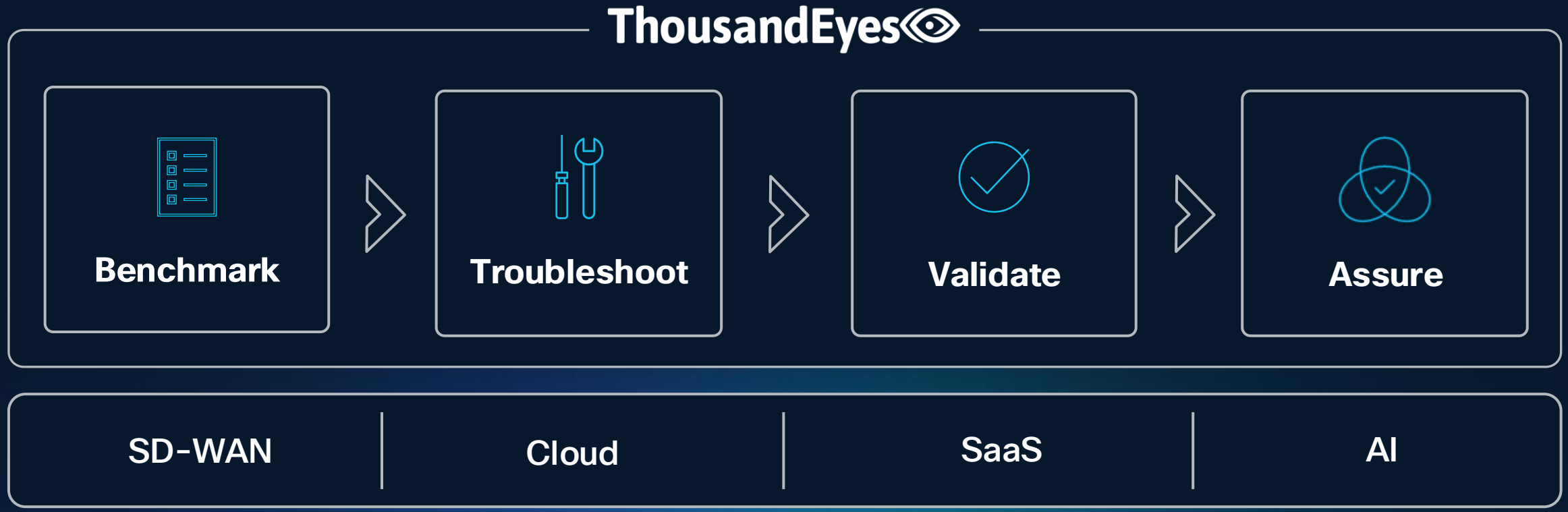
Shawn Eustis
Director Solutions Engineering

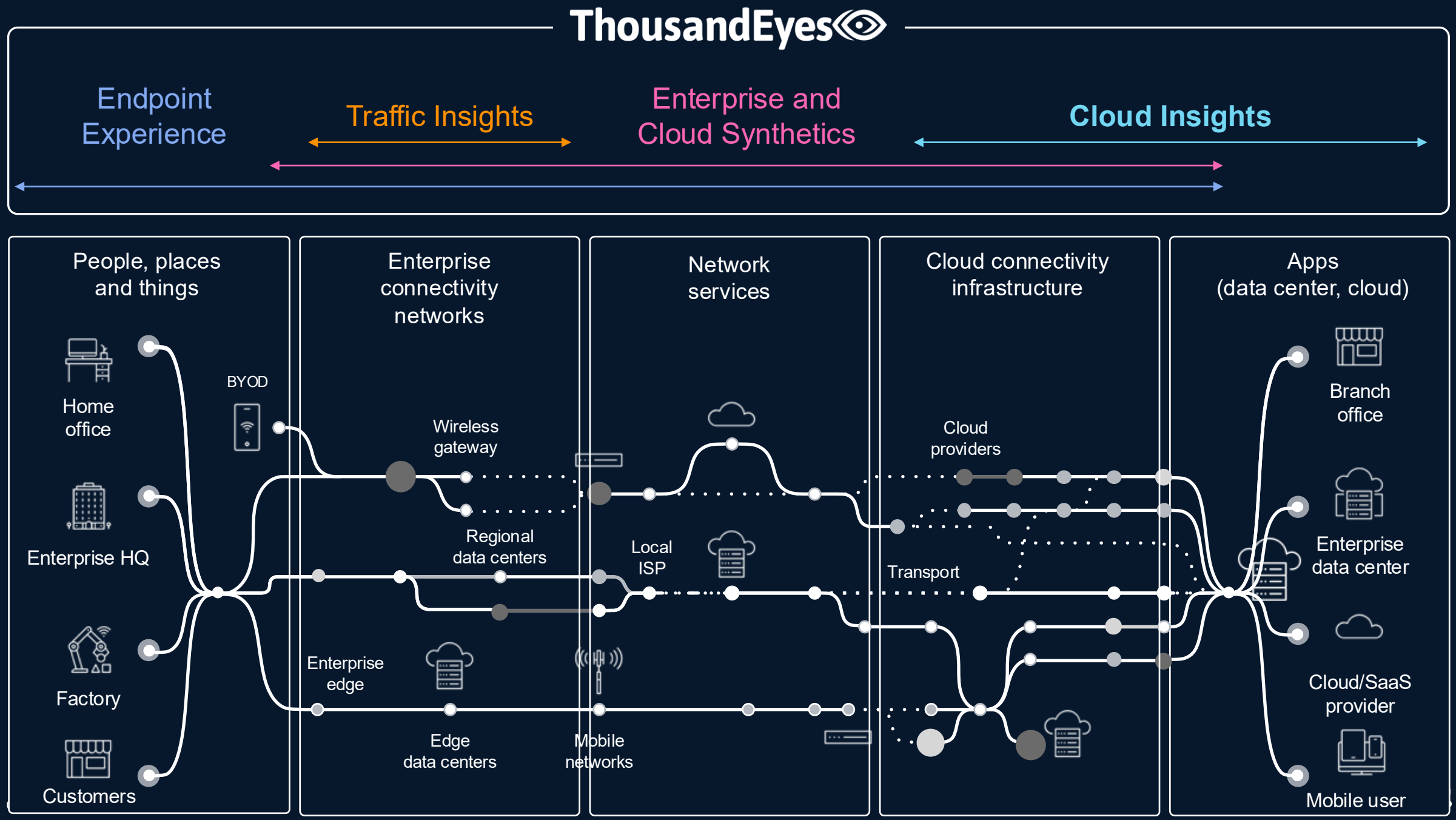
March. 18th, 2026

Vacation Vibe Coding

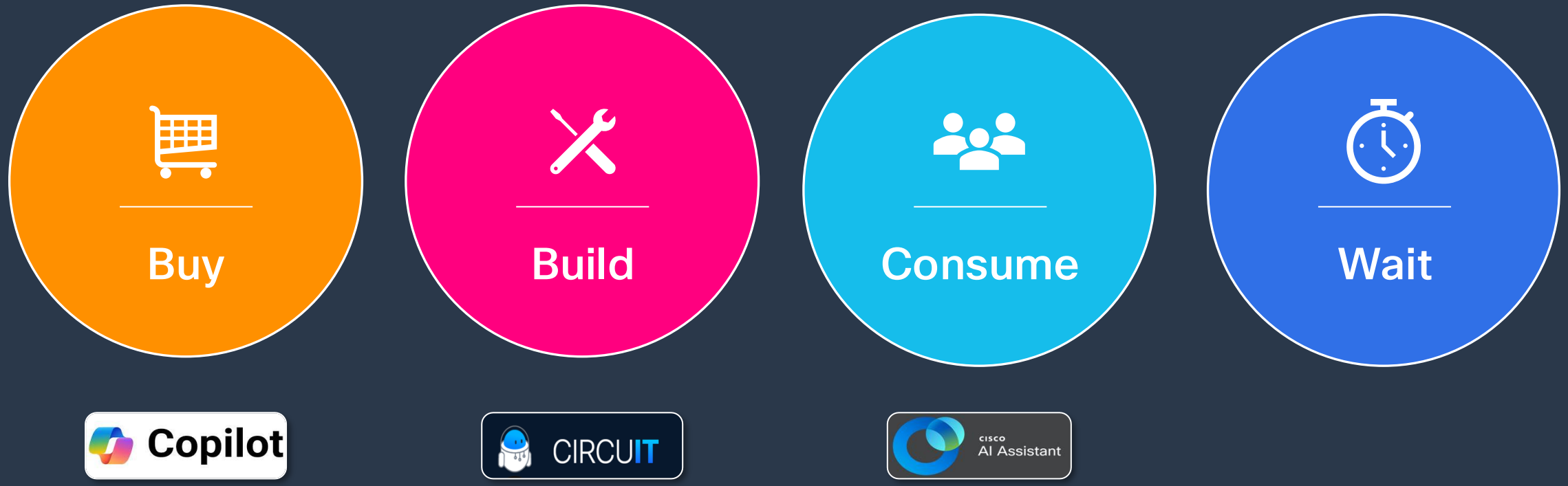
De-risk and Accelerate Transformative Projects

Actionable Insights Before, During, & After




ThousandEyes
Endpoint Experience
Traffic Insights
Enterprise and Cloud Synthetics
Cloud Insights
People, places and things
Enterprise connectivity networks
Network services
Cloud connectivity infrastructure
Apps (data center, cloud)
Home office
BYOD
Enterprise HQ
Factory
Customers
Wireless gateway
Regional data centers
Enterprise edge
Edge data centers
Local ISP
Mobile networks
Cloud providers
Transport
Branch office
Enterprise data center
Cloud/SaaS provider
Mobile user

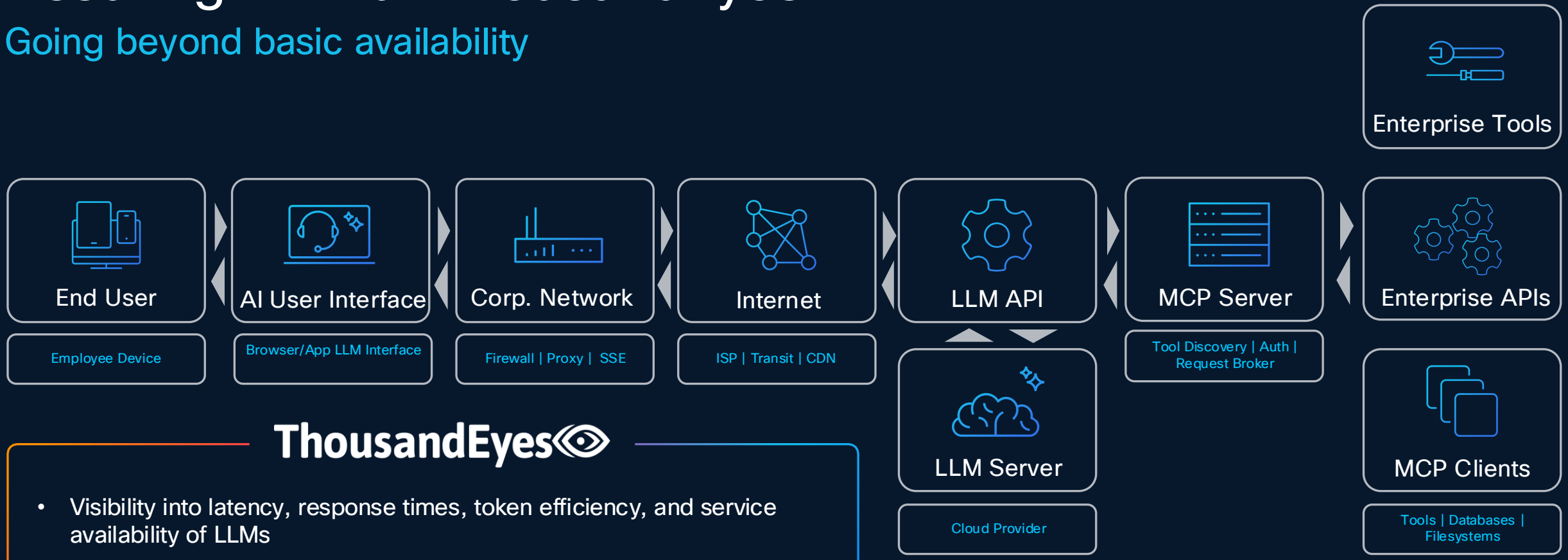
Where are you on your AI journey?



Why Assurance?

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

cisco

72 hours.

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

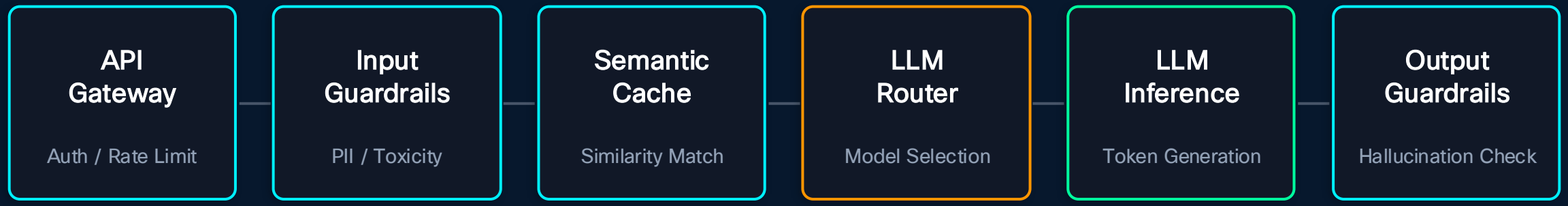
The model didn't go down. The infrastructure didn't fail.

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

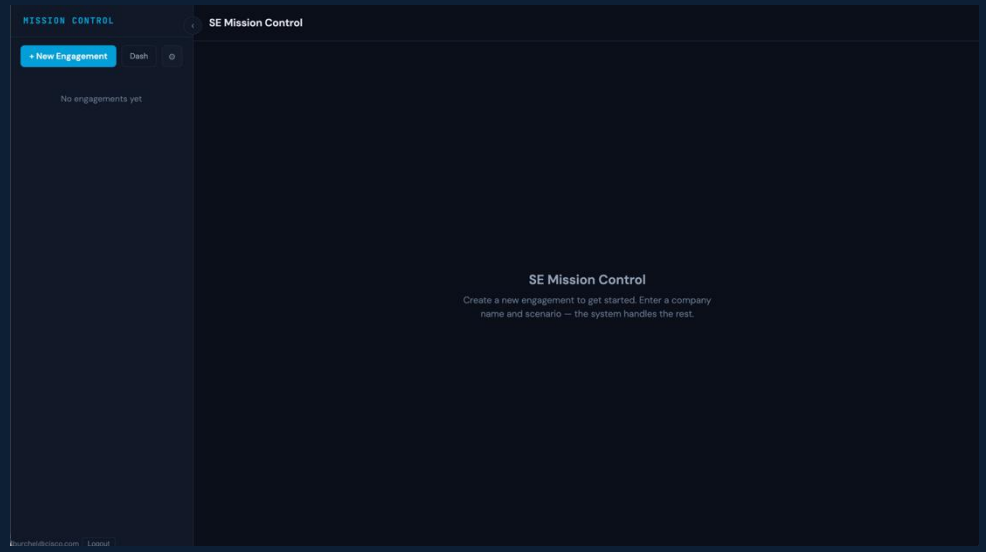
60%
Of query latency occurs outside the LLM inference later

We built an AI Application.

Mission Control

se-missioncontrol.com

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

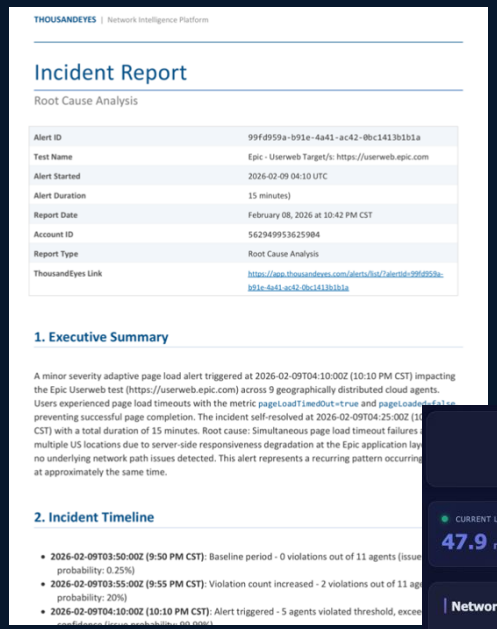
- Internet Insights: OpenAI + Duo outage detection
- Endpoint agent: SE laptop → Mission Control path
- Alert → Webex notification
- Weekly AI dependency health summary

Demo

ThousandEyes AI Assurance

Agentic AIOps – MCP Server

ThousandEyes MCP Server

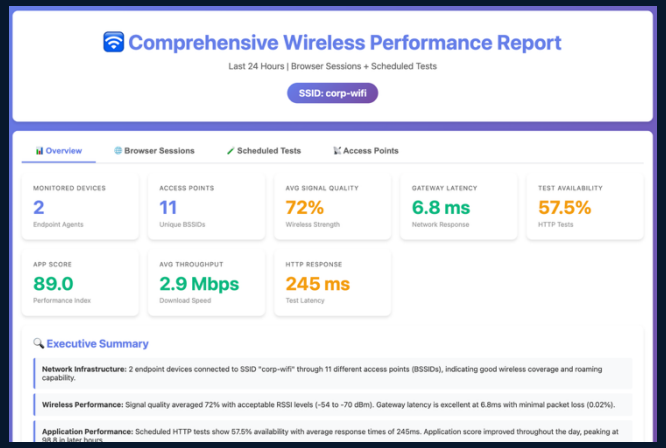


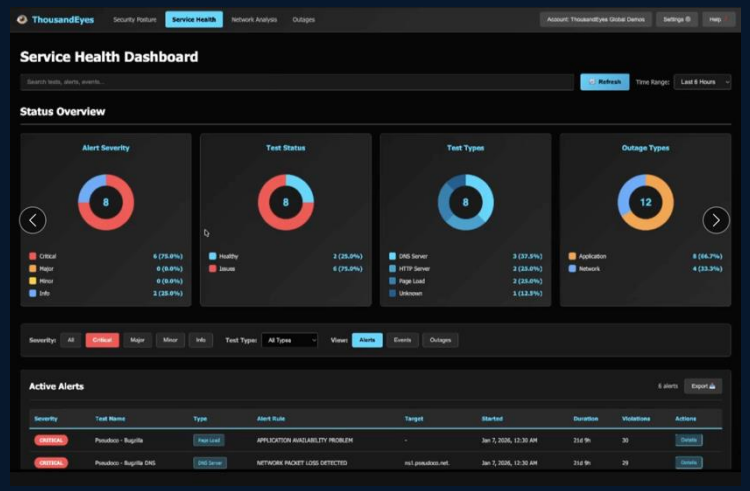
Done. I've created the incident:

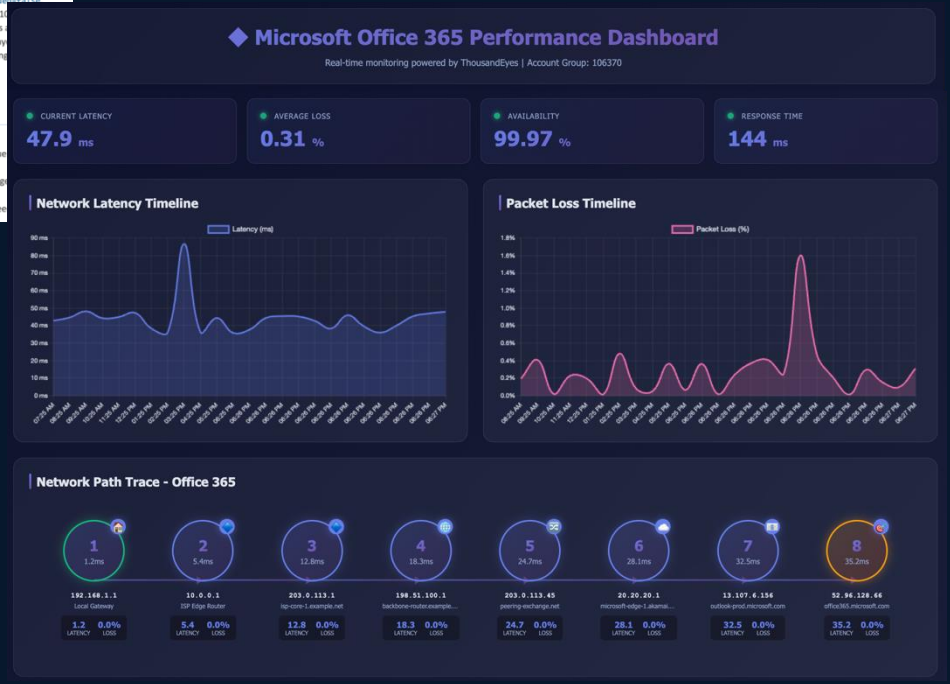
Incident Number: INC0080991

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,



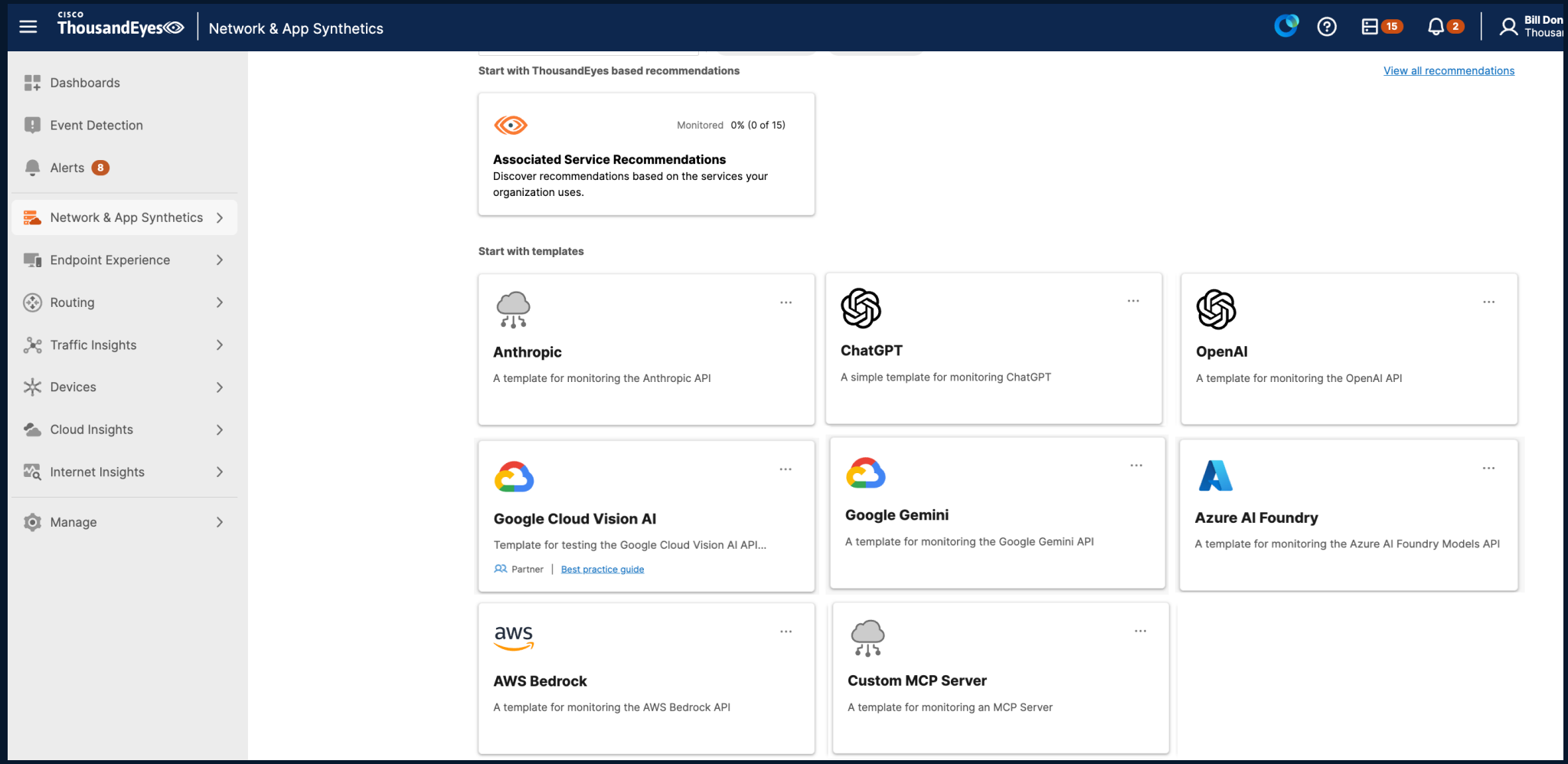




AI Assurance Demos

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.	https://ciwlpuvdopteilegxzaszwcixfxquiwz.share2.thousandeyes.com
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.	https://crzleyggtjvfcbyawetlmljvikltgzmq.share2.thousandeyes.com
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.	https://auvoutttttlkrsfwzkgpxcphpfojimhfa.share.thousandeyes.com
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.	https://ctavzhxqdnyaiqdbjglucnnjllavcjvm.share2.thousandeyes.com

You Can Get Started Now



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.


CISCO