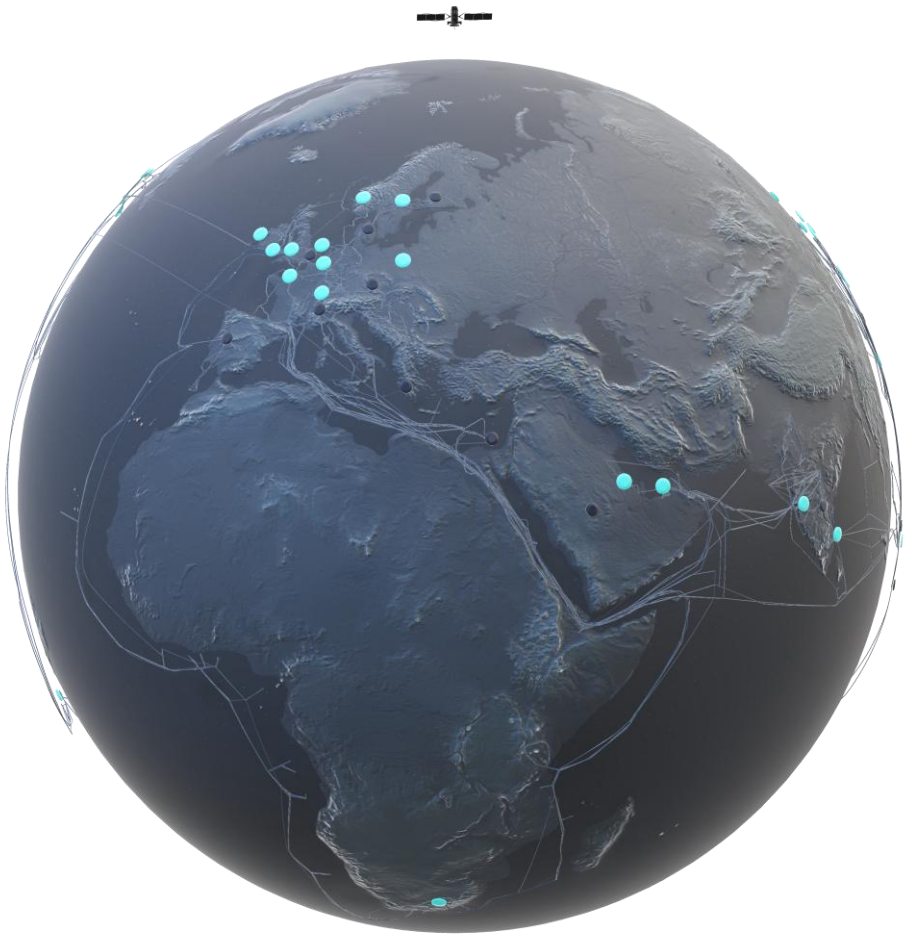
AIOps Unleashed: Transforming Cloud Operations at Microsoft Azure

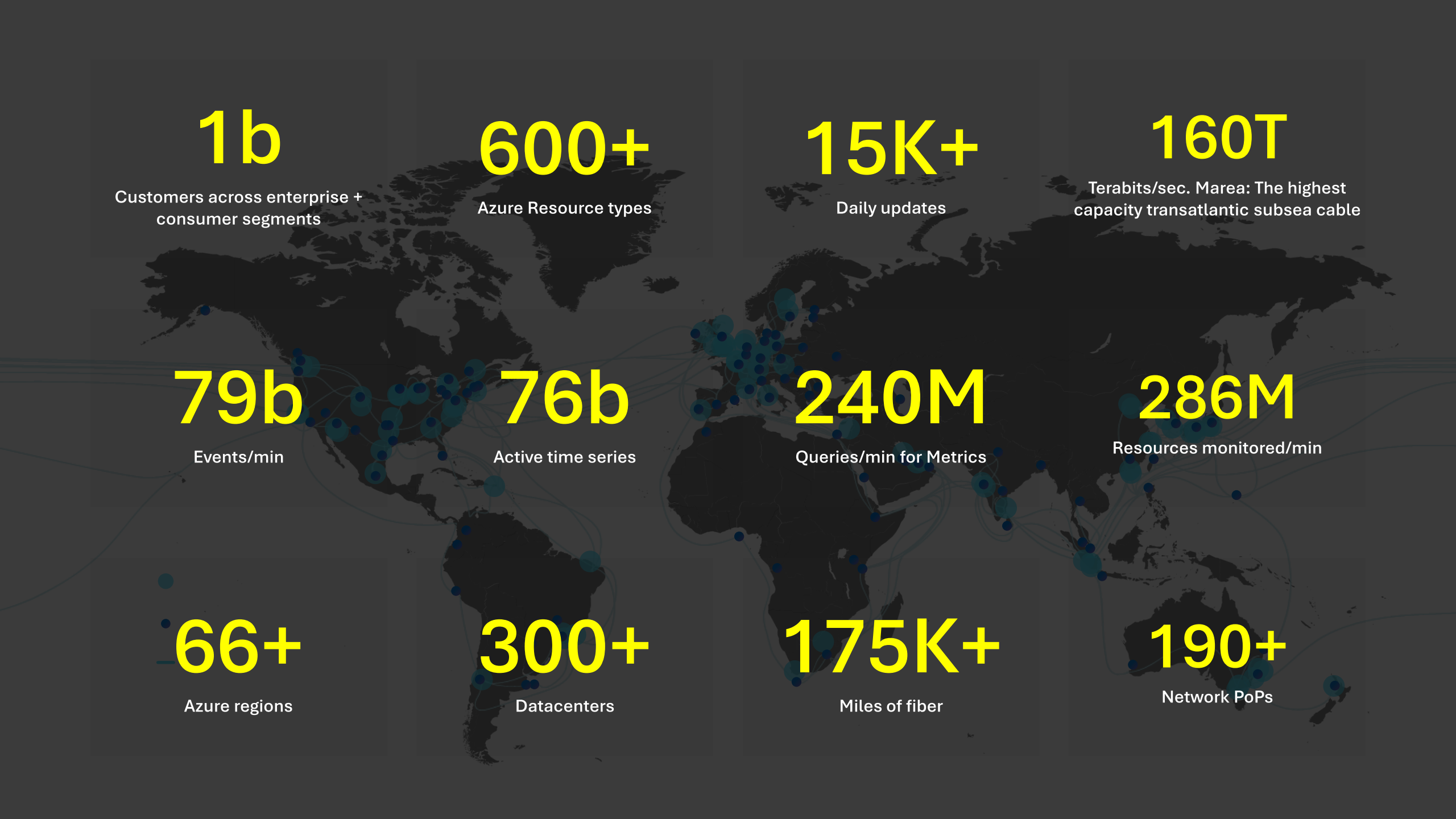
Zhangwei Xu
VP/Distinguished Engineer, Microsoft

Azure

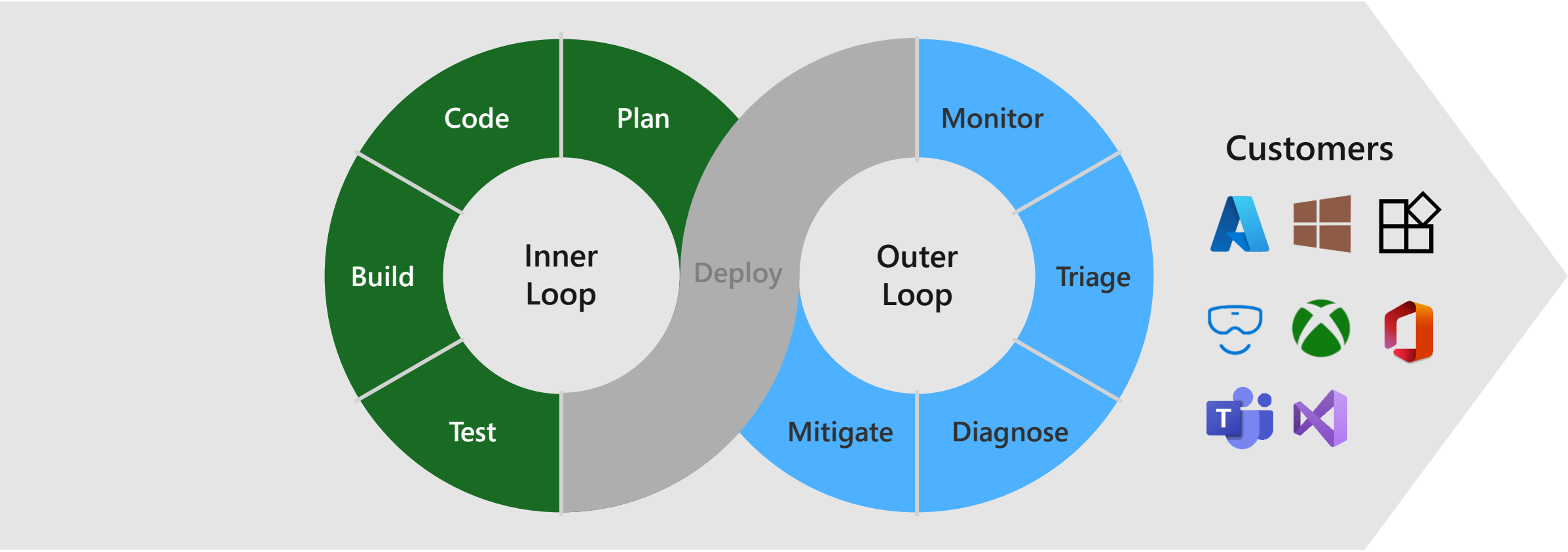


Azure

The world's
AI supercomputer


1b
Customers across enterprise + consumer segments
600+
Azure Resource types
15K+
Daily updates
160T
Terabits/sec. Marea: The highest capacity transatlantic subsea cable
79b
Events/min
76b
Active time series
240M
Queries/min for Metrics
286M
Resources monitored/min
66+
Azure regions
300+
Datacenters
175K+
Miles of fiber
190+
Network PoPs

Building and Operating Azure



Operating Microsoft Clouds – The tools

- Build Systems
- CI/CD Pipelines/Test platforms
- Dev/Service Environments
- Documentations
- Work Item Management tool
- Deployment Systems
- Change Management Systems
- Safe/Secure Production Touch systems
- Automation systems
- Security/Auditing systems
- Telemetry Platforms
- Monitoring Systems
- Inventory/Graphs
- Analytics platforms
- Incident Management System
- Diagnostic tools
- AIOps Systems

Challenges Facing Service Teams, Operator

Current state of the art monitoring systems

Handles large volume of telemetry data at scale with low latency

Provides a scalable platform to create targeted monitors & alerts

Provides incredible flexibility and sophistication to do data analysis & alerting

Today's systems require humans with intricate knowledge & expertise who understand the Service and the monitoring systems

Challenges

Visibility

Understand what's going on at any given time

- Limited holistic health view
- Needs human interpretation for numerous metrics/logs
- Prone to alert storms

Scale & Efficiency

Global infrastructure and service(s) + High volume of tickets

- Increasing number of monitors contributes to high noise (for Oncall engineers)
- Need to keep up with large volume of telemetry
- Need to manually manage monitors
- Rule based monitors are usually tricky to tune

Complexity

Fragmented and intricately connected services

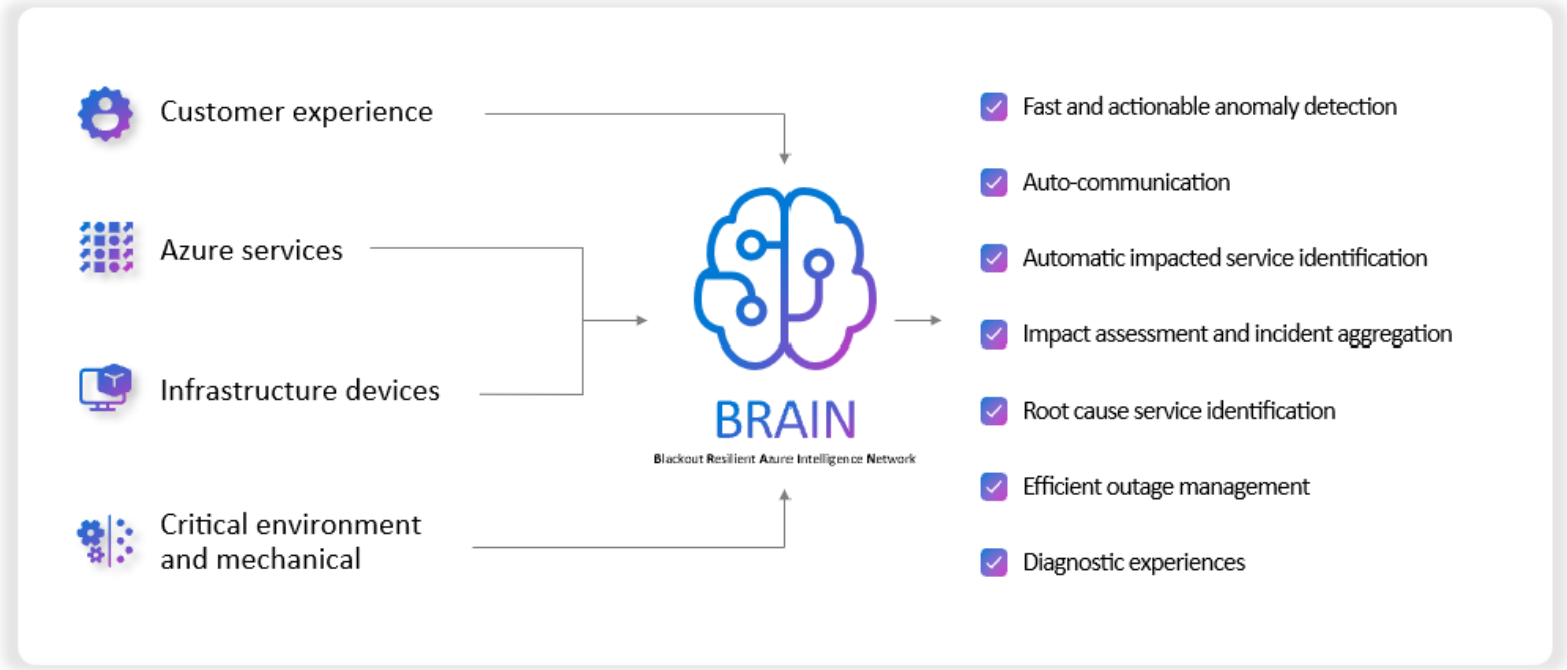
- Require complex configuration, tuning and dashboarding skills
- Must manually create automation tasks & workflows
- Depends on human intelligence for quality assessment

Azure AIOps Vision & Strategy

Significantly improve service reliability by preventing outages, minimize impact, and eliminate human toil

For 1P and 3P customers

1. **Standardize and Centralize**
   Standard signals (SLI), centralized digital twin, system wide health & monitoring (BRAIN)
2. **Automate operations**
   Evaluate, assess, detect, troubleshoot, mitigate
3. **Shift-left**
   Prevent at source, Predict & mitigate before issues snowballs
4. **Leverage AI & ML**
   Learn from changing conditions & handle complex environments
5. **Extensible platform**
   Extensible ingestion & consumption experiences



BRAIN Components

Health

Holistic health

Accurate health signals incorporating all signals

Provide health signals for all other systems

AI Monitor

Monitors all telemetry

High precision/recall

Auto tuning

Auto configuration

Auto correlation

Multi-dimensional Anomaly Detection

Seamless Integration with traditional monitors

AI Automation

Auto outage declaration

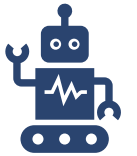
Auto customer communication

Auto Triage

Auto RCA

Auto Mitigation

Auto Deployment Health Check



Intelligent Engine

Self learning using all data sources

Uber DRI

Auto-TSG

Copilot experience

AI Monitors

Traditional monitors

AI monitors

Traditional monitor	Area	AI Monitor
Manual	Setup	Automatically monitor all telemetry
Manual	Management	Auto managed
Reactive, Static rule	Trigger	Predictive, Dynamic, Adaptive
Fixed telemetry/dimension/query	Monitor approach	Multiple telemetry/dimension simultaneously
Manual	Tuning	Automatic
No	Learning	Continuous self learning based on feedback
Limited	Insights	Comprehensive health and alert enrichment
Limited or require other systems	Correlation	Built-in auto correlation
Manually configured	Action	Built-in intelligent automation

BRAIN benefits and results

68%

Time
to outage
reduction

45%

Time
to mitigation
reduction

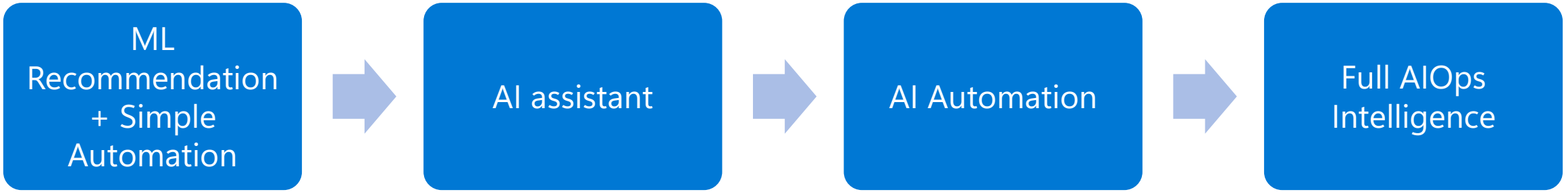
58%

Time
to notification
reduction

25%

Incident
noise
reduction

AIOps Journey – War of Control, from ML to AI



- Human engineers doing all work
- Review recommendations
- Pre-configured simple automation

- Human engineers doing most work
- More dynamic automations
- Replacing some rules

- Automating human actions powered by GenAI
- Based on human knowledge
- Operating human written code
- Generic rules engine
- Specialized agents

- AI operating AI written code
- Human engineers overseeing
- Self learning and evolving agents covering entire product lifecycle

Human and AI Collaboration

- Building trust
  - Confidence erodes quickly and no single FP/failure accepted
- Control
  - Tendency to apply static rules which defeat the purpose
  - Human override
- Explanation
  - Need to understand why for everything
- Manual + Automation + AI
  - Hybrid operation mode is the reality before full AIOps is achieved
- Knowledge extraction and learning
  - Tribal knowledge

Future of AIOps

- Many of the Ops problems today can be attributed to human
  - Changes, bugs, not following process, insufficient capacity
- When AI writes all the code, will there still be Ops as we know today?
  - Self-operating code and systems
  - Shift left to solve operation problems from the root
- Would AI create framework and principles?
  - Guiding principles
- Auto-* - Auto-scale, Auto roll-back, Auto update, Auto migration, Auto health/validation, Auto-monitoring