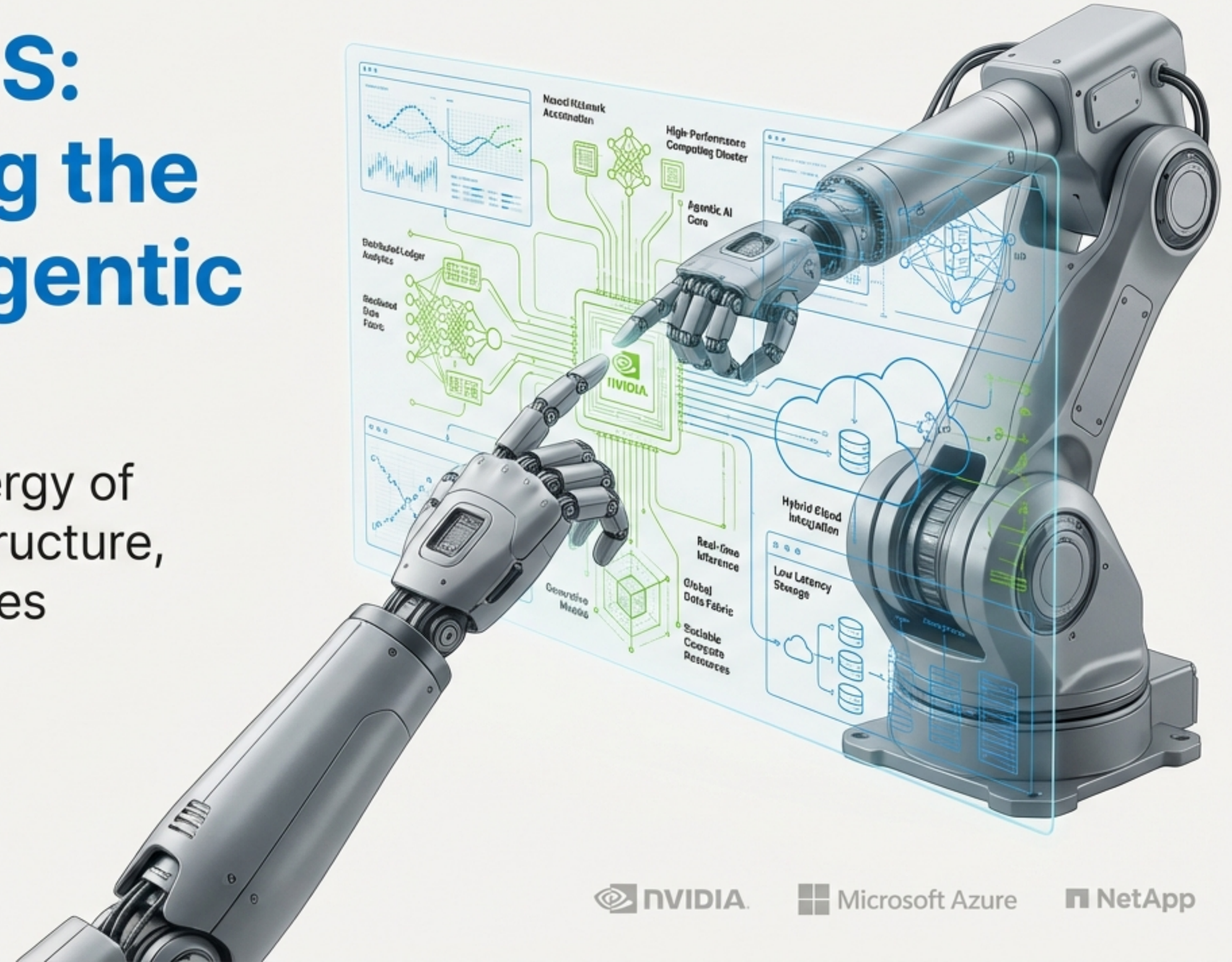
Ascend_EOS: Architecting the Future of Agentic Agentic AI

The Unmatched Synergy of NVIDIA, Azure Infrastructure, and Azure NetApp Files



NVIDIA Microsoft Azure NetApp

The New Frontier: From Predictive Models to Interactive AI Agents

The industry is moving beyond static AI. The future is 'Physical AI'—embodied agents that perceive, reason, and act in complex environments.

These agents require more than just pattern recognition; they need **contextual, step-by-step decision-making.**

This shift unlocks transformative use cases: real-time robotics, autonomous factories, intelligent digital twins, and smarter city infrastructure.

Key Terms

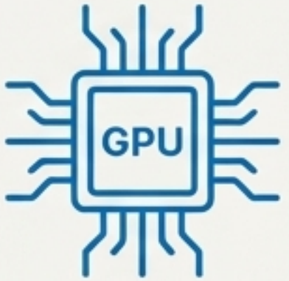
- Physical AI
- Embodied agents
- Real-time robotics
- Contextual, step-by-step decision-making

The Core Challenges of Production-Grade Agentic Systems

Building robust AI agents is a multi-faceted engineering challenge that legacy infrastructure cannot solve. It demands a purpose-built, integrated stack.

1. Extreme Data Throughput

Constant access to massive datasets for training and sub-millisecond latency for real-time inference.



2. Massive, Scalable Compute

Purpose-built infrastructure to train and run state-of-the-art models efficiently across distributed environments.

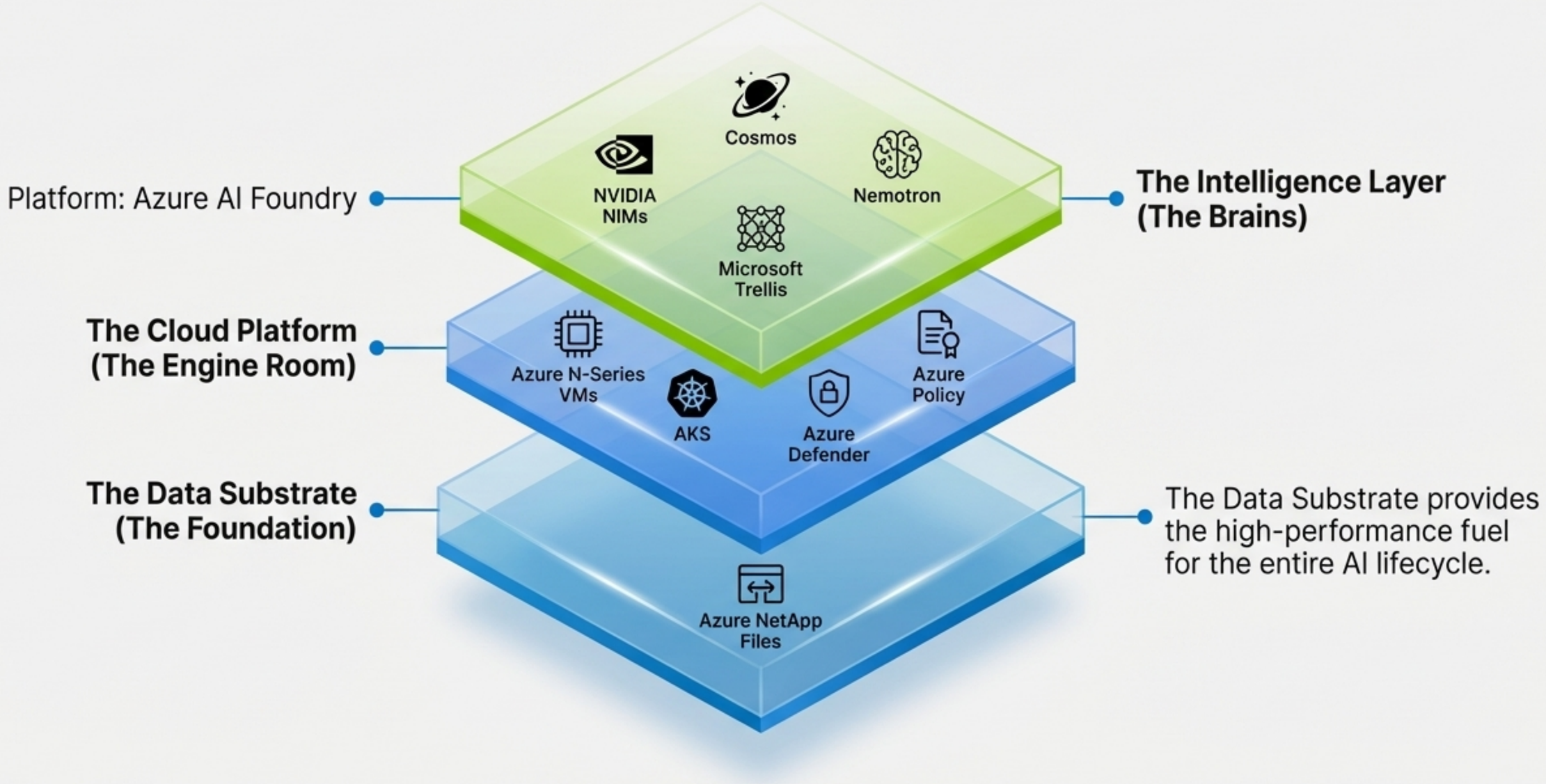
3. State-of-the-Art Intelligence

Access to models that go beyond language to perform complex reasoning about the physical world.

4. Enterprise-Grade Security & Governance

The ability to deploy these powerful systems safely, with robust threat protection, compliance guardrails, and operational control.

A Blueprint for Agentic AI: The Ascend_EOS Architecture



The Foundation: Fueling the AI Engine with Azure NetApp Files

AI agents are data-intensive. Their performance is directly tied to the speed and scalability of the underlying data platform.

Extreme Performance

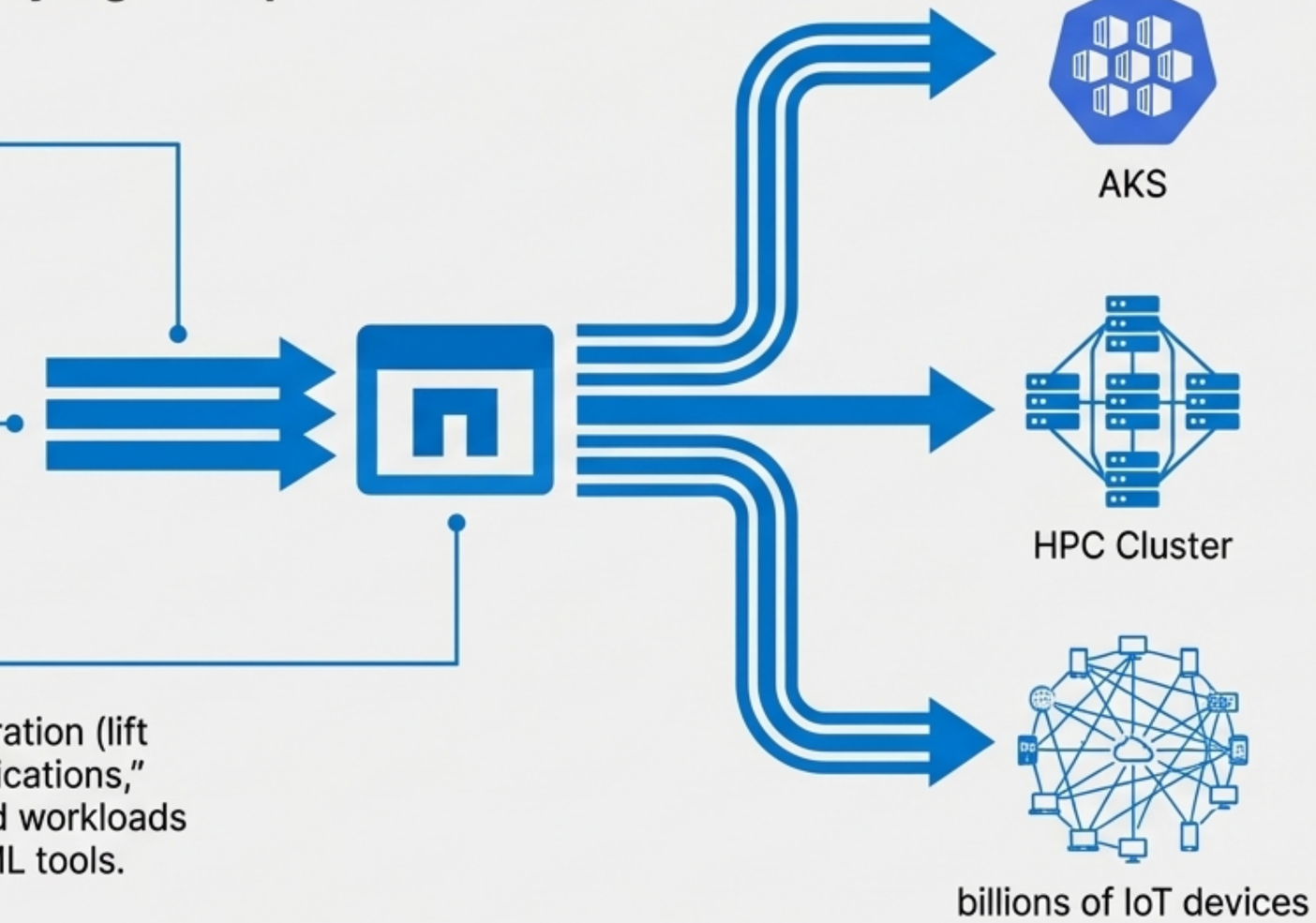
Delivers “consistent sub-millisecond latency” with high IOPS and throughput, essential for parallel processing with tools like NVIDIA RAPIDS and Dask.

Seamless Scalability

Scales to meet the demands of both HPC applications and data collection from “billions of IoT devices.”

Natively Integrated & Versatile

Designed for mission-critical workloads including “migration (lift and shift) of POSIX-compliant Linux and Windows applications,” databases, and as persistent volumes for containerized workloads on AKS. The POSIX compliance is critical for many AI/ML tools.



The Engine Room: Scalable, Secure, and Governed AI Infrastructure on Azure

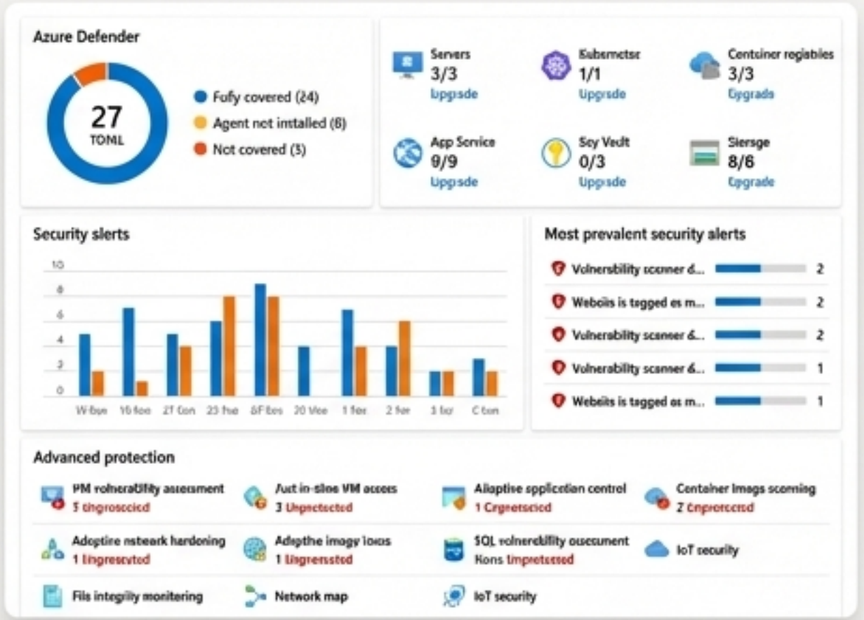
Purpose-Built Compute

GPU-based **N-Series Virtual Machines** for accelerated training and inference.

CPU-based **H-Series Virtual Machines** for demanding HPC workloads.

Azure Kubernetes Service (AKS) & **Virtual Machine Scale Sets** for orchestrating and scaling distributed jobs at massive scale.

Comprehensive Security & Governance



Azure Defender provides advanced threat protection for hybrid environments.

Azure Policy establishes 'guardrails' to enforce organizational standards.

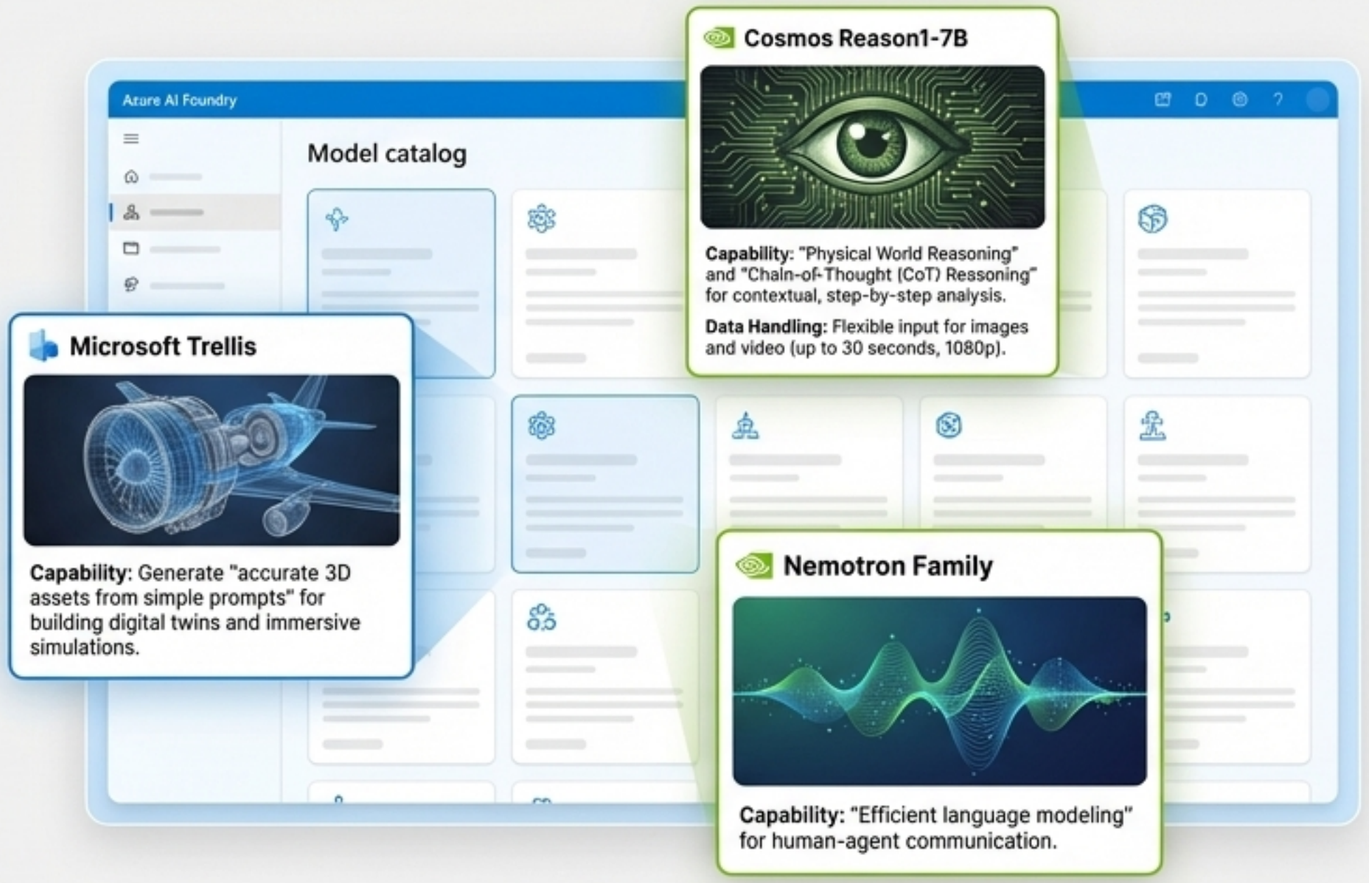
A **Zero Trust Model** provides a foundational security strategy.

Operational Excellence

Built upon the best practices of the **Microsoft Azure Well-Architected Framework**, covering the five pillars of reliability, security, cost optimization, operational excellence, and performance efficiency.

The Brains: Perception and Reasoning with NVIDIA NIMs
Inter SemiBold

State-of-the-art NVIDIA models, delivered as **optimized microservices**, provide the intelligence for agents to understand and **interact with the physical world.**
Inter Regular



Deployment Platform: These models are available as **NVIDIA NIM (NVIDIA Inference Microservices)** in the **Azure AI Foundry**, enabling secure, scalable deployment in minutes.

The Ascend_EOS Flywheel: A Virtuous Cycle of AI Innovation

DATA PREP

Massive, unstructured data (video, sensor) is ingested and prepared on **Azure NetApp Files**, accelerated by GPU-powered libraries like **NVIDIA RAPIDS**.

DISTRIBUTED TRAINING

Models are trained at scale on **Azure N-Series VMs** within an **AKS cluster**, using frameworks like **Horovod** to pull data directly from ANF persistent volumes.

OPTIMIZED DEPLOYMENT

State-of-the-art models are packaged as **NVIDIA NIMs** and deployed securely and efficiently through the **Azure AI Foundry**.

AGENTIC INFERENCE

Deployed agents perform complex reasoning (**Cosmos Reason1-7B**) and interact with digital twins (**Microsoft Trellis**), generating new data and insights.

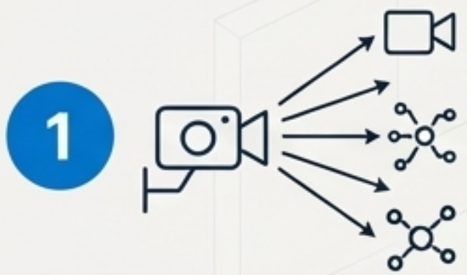
FEEDBACK LOOP

The new data is written back to **Azure NetApp Files**, restarting the cycle for continuous learning and model refinement.

Use Case in Action: Autonomous Factory Floor Monitoring

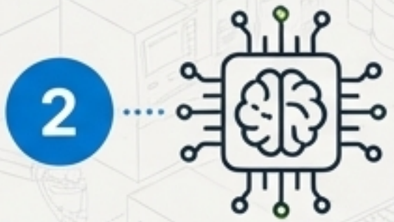
Scenario Overview: An AI agent monitors a factory for safety compliance and predictive maintenance.

1. PERCEIVE



'Millions of live streams' and sensor data are ingested and stored on **Azure NetApp Files**, providing the low-latency access required for real-time analysis.

2. REASON

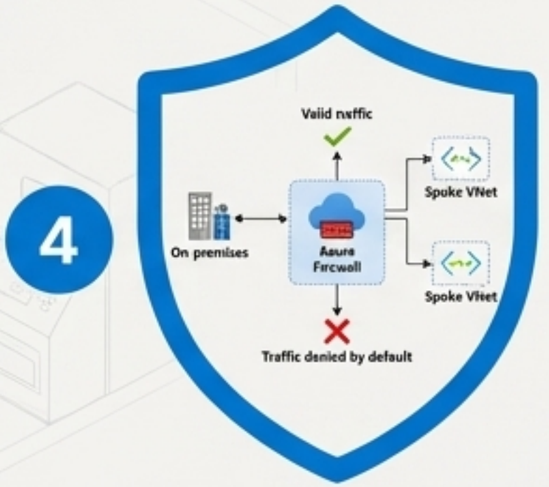


A **Cosmos Reason1-7B NIM**, running on an **AKS cluster**, uses **Chain-of-Thought** reasoning to analyze video feeds, identify a potential safety hazard, and predict an impending machine failure.

3. ACT

The agent automatically triggers a safety alert, updates a **Microsoft Trellis**-powered digital twin of the factory floor, and dispatches an autonomous maintenance robot to the location.

4. SECURE & GOVERN



The entire workflow is protected by **Azure Defender**'s threat detection, and resource configurations are enforced by **Azure Policy** to maintain security and compliance standards.

A Unified Platform for Enterprise-Grade Agentic AI

Unrivaled Performance at Every Layer

The combination of sub-millisecond data access with Azure NetApp Files (highlighted in #0067C5) and purpose-built AI compute on Azure N-Series GPUs (highlighted in #0078D4) eliminates bottlenecks across the entire AI lifecycle.

Cutting-Edge Intelligence, Simplified

Access state-of-the-art NVIDIA models like Cosmos #76B900 for physical world reasoning, delivered as production-ready NIM microservices #76B900) on the managed Azure AI Foundry #0078D4) platform.

Enterprise-Ready from Day One

Deploy with confidence on a foundation of comprehensive security (Azure Defender, #0078D4), robust governance (Azure Policy, #0078D4), and operational best practices defined by the Azure Well-Architected Framework.

Build the Future of Agentic AI, Today

The tools to build the next generation of AI are available now.
Here is how you can begin.

1.

EXPLORE

Browse the complete model catalog, including Cosmos, Nemotron, and Trellis, in the **Azure AI Foundry**.

2.

DEPLOY

Use the one-click deployment for **NVIDIA NIM Microservices** to stand up a proof-of-concept in minutes.

3.

ORCHESTRATE

Leverage the **NVIDIA NeMo Agent Toolkit** to build, monitor, and optimize your collaborative AI agents.

Scan to explore
Azure AI Foundry