World Wide Technology

Make a new world happen

WWT's AI Proving Ground

Miguel Mateo - Chief Technology Advisor

June 2024

WWT’s AI Proving Ground Lab

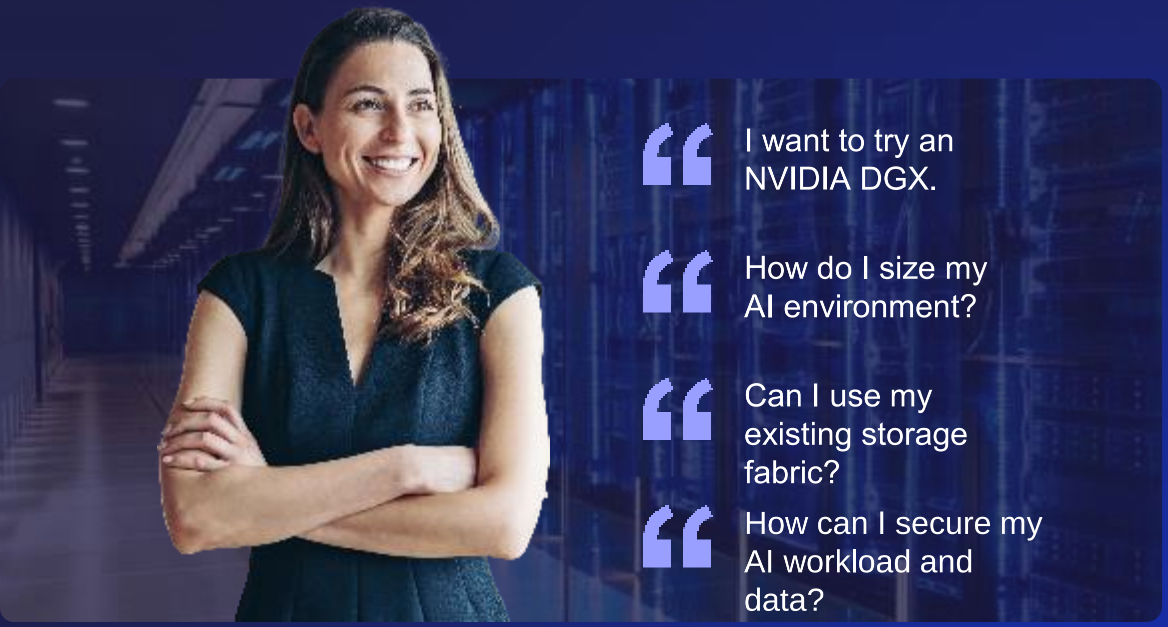
The industry’s first and most comprehensive AI testing environment

AI ecosystem enablement

Generative AI and deep learning

Edge compute and AI inference

Foundational data capabilities



“ I want to try an NVIDIA DGX.

“ How do I size my AI environment?

“ Can I use my existing storage fabric?

“ How can I secure my AI workload and data?

intel · NetApp · DELL Technologies · NVIDIA · CISCO

AMD · IBM · Hewlett Packard Enterprise · ARISTA · PURESTORAGE

VAST · TensorFlow · mlflow

kubernetes · kafka · Google Cloud · Azure · aws

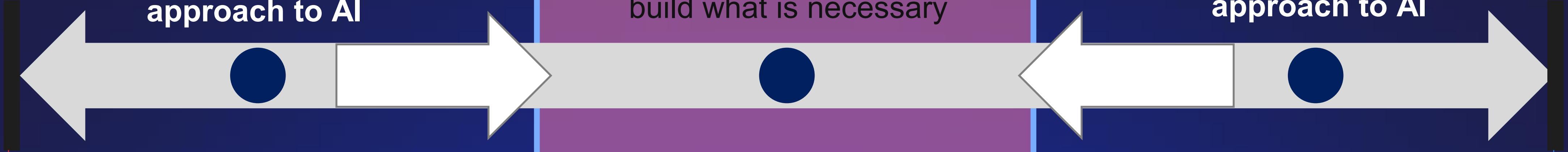
Infrastructure · Software · Public cloud

High-performance compute • Storage for AI • Memory for AI • High-speed networking • Security and governance • Data pipelines

Testing frameworks • Cluster management • Version control • Deployment API • LLM library • IaaS vs. PaaS • Hybrid frameworks

Balancing actionable outcomes with a scalable long-term strategy

We cut through the hype and build practical solutions for customers



IMPACT	Off-the-shelf "software integration" approach to AI	**Practical AI** **Prioritize precision and speed-to-outcome:** buy where available, build what is necessary	Hyper-custom R&D approach to AI
Short-term	• High initial cost to license • Difficult to customize • Fast adoption and training • Scalable • POC strain	• Promotes innovation and speed • Security and governance embedded from incubation • Optimizes cost and scalability • Directly drives business outcomes	• Experimental architectures that are hard to scale • Feature adoption and challenges • Long time to market
Long-term	• AI gets decentralized and harder to manage / govern • Difficult to gain competitive edge	• AI value is perceived cohesively across organization • Strengthens regulatory posture through responsible AI, ethics and governance	• Very costly • Solutions perceived as experimental and not ready for primetime • Highly customized proprietary solution to fit exact business need

Three key building blocks for HPA

COMPUTE

NVIDIA AMD intel

- HPC / supercomputing
- Accelerated computing
- Heterogenous computing
- Emergent computing
- Quantum computing

STORAGE

DELL PURE STORAGE IBM

- Parallel file system storage
- Streaming storage
- Synthetic data
- Computational storage
- Emergent storage

NETWORK

NVIDIA CISCO ARISTA

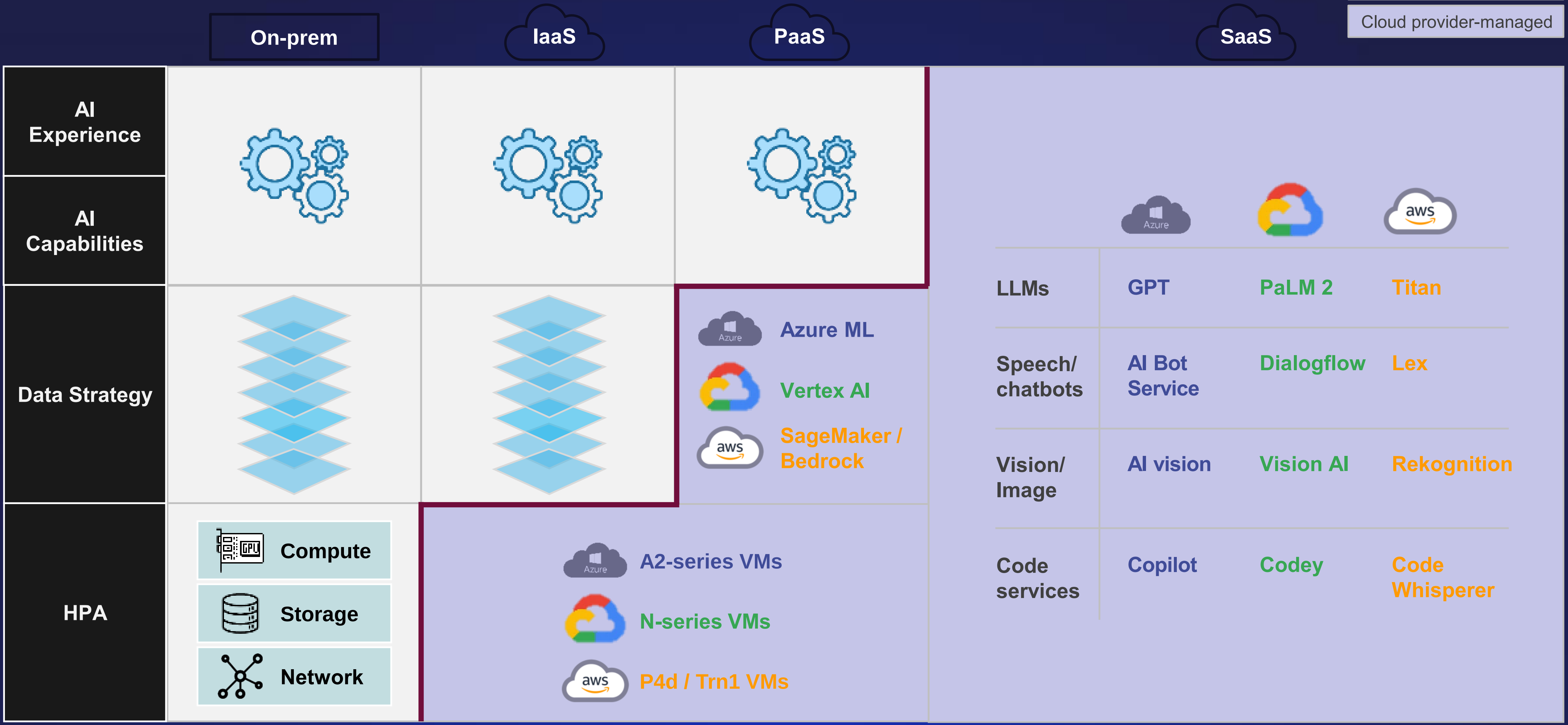
- Connects users and infrastructure
- Secure, smart, fast fabrics
- SmartNICs and data processing units
- Computational networking
- Photonics (SOC, switches, backplanes)

Financial ROI and Innovation ROR Results from Investments in HPC: HPC ROI can reach $507 dollars in sales revenues per dollar invested, and $47 dollars in profits or cost savings per dollar invested in dedicated strategic HPC activities. – *Hyperion Research*

Various consumption models can be built on-prem or in the cloud, offering multiple options for system management

Customer-managed

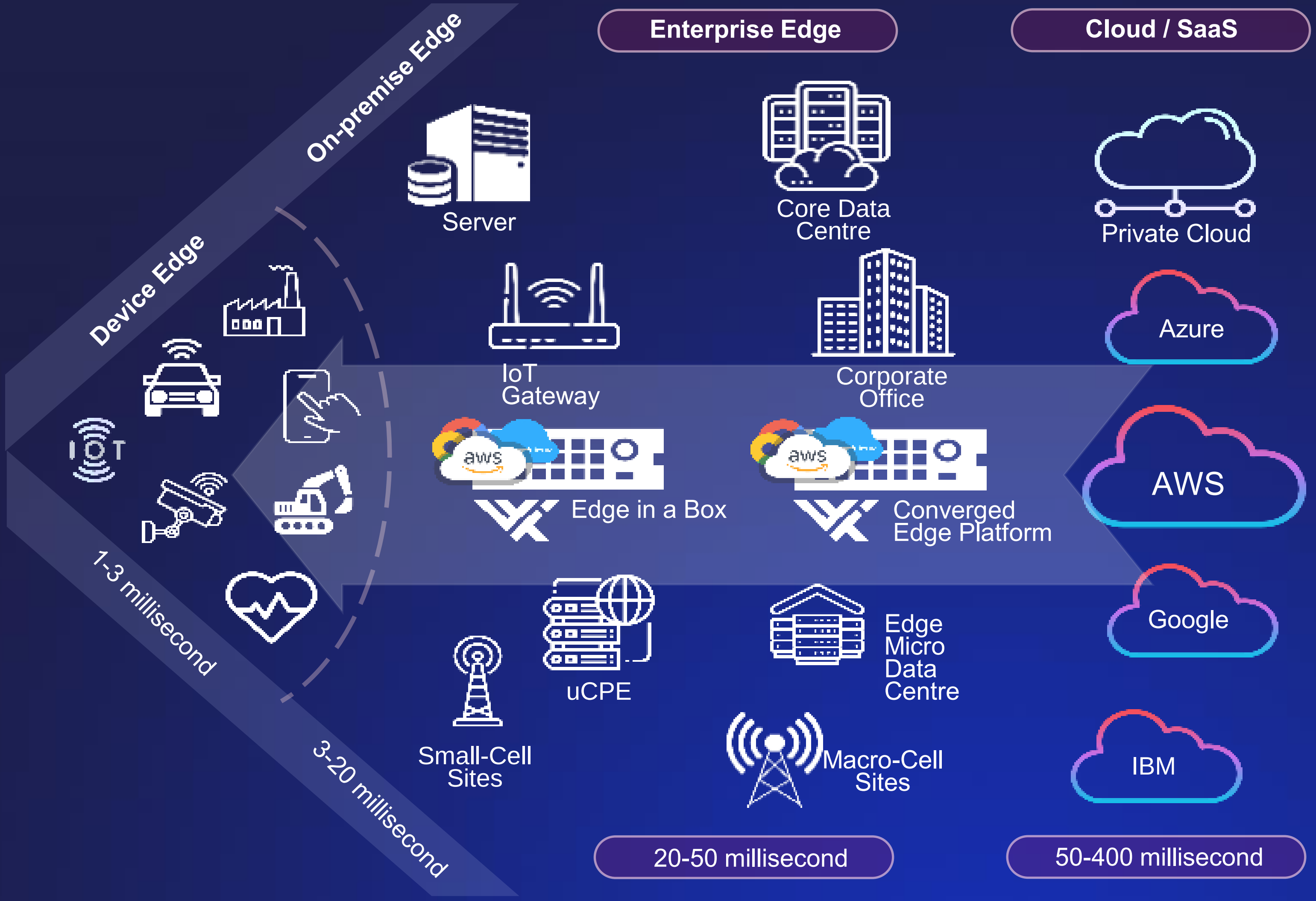
Cloud provider-managed



	On-prem	IaaS	PaaS	SaaS
AI Experience				
AI Capabilities				
Data Strategy			Azure ML Vertex AI SageMaker / Bedrock	
HPA	Compute Storage Network	A2-series VMs N-series VMs P4d / Trn1 VMs		

	Azure		aws
LLMs	GPT	PaLM 2	Titan
Speech/ chatbots	AI Bot Service	Dialogflow	Lex
Vision/ Image	AI vision	Vision AI	Rekognition
Code services	Copilot	Codey	Code Whisperer

Edge devices can provide lower latency for AI use cases



KEY USE CASES

- **IoT & Smart Devices** - These devices can run applications on their own, rather than in the cloud, to deliver highly responsive and efficient user interactions
- **Retail Experience** - Stores can deliver a personalized experience enriched with product recommendations, instant assistance, and digital promotion to bridge the online and offline customer experience
- **Video Conferencing** - Streaming interactive live video takes up a lot of bandwidth. Moving backend processing closer to the data source can help decrease lag and latency
- **Self Driving Cars** - Edge enables autonomous vehicles to react in real time instead of waiting for information to be transferred to and from a server
- **Security Monitoring** - High-definition IoT video cameras can analyze footage before sending relevant information to a cloud server for further action to reduce bandwidth usage
- **Medical Monitoring** - Edge computing capabilities enable these devices to respond in real time without having to transmit the health data to a cloud server and losing time

Introducing The AI Proving Ground

ATC

WWT's ATC is building the industry's first and only multi-OEM AI testing ground

A state-of-the-art environment where product developers can replicate their AI workloads
Experiment, test and innovate with the latest from the world of AI in a secure and scalable manner

Lab components and OEMs

INFRASTRUCTURE

intel · Hewlett Packard Enterprise · DELL · ARISTA · NVIDIA · AMD · PURESTORAGE · VAST · NetApp

- ✓ High- performance compute
- ✓ Storage for AI
- ✓ Memory for AI
- ✓ High speed networking

MIDDLEWARE

kafka · mlflow · TensorFlow · GitHub · kubernetes · Prometheus

- ✓ Security and governance
- ✓ Data pipelines
- ✓ Testing frameworks
- ✓ Cluster management
- ✓ Version control
- ✓ Deployment API

PUBLIC CLOUD

Microsoft Azure · aws · Google Cloud

- ✓ LLMs Library *OpenAI, Meta, …*
- ✓ IaaS vs. PaaS
- ✓ Hybrid frameworks

Level of customization

On demand
Demo, sandbox, training environment

Proof of concept
Single test plan and success criteria

Lab as a service
Dedicated, long-term customer lab

Lab services

AI ecosystem enablement

- Thermal modeling and ESG impact estimation
- GPU capacity forecasting and right-sizing
- AI stack comparisons (e.g., InfiniBand vs. Ultra-Ethernet)
- Public cloud vs. Specialist GPU cloud vs. on-prem bake-offs
- TCO estimation for SaaS vs. custom AI products

Generative AI and deep-learning

- LLM fine-tuning (cloud and on-prem)
- Computer vision and image modeling
- Vector DBs selection and LLMOps

Edge-compute and AI inference

- Edge frameworks and AI inference
- Testing LLM/GenAI embeddings in edge-compute products

Foundational data capabilities

- Digital twins, AI workload replication
- Federated machine learning
- AI middleware: data catalogs, lineage tools, etc.

World Wide Technology
Make a new world happen

WWT’s AI Proving Ground Lab

The industry’s first and most comprehensive AI testing environment

Thank you!