



Forgiare l'AI

Soluzioni a confronto dal Cloud all'Edge con Azure Foundry



Marco Dal Pino

Senior Technical Consultant, Microsoft



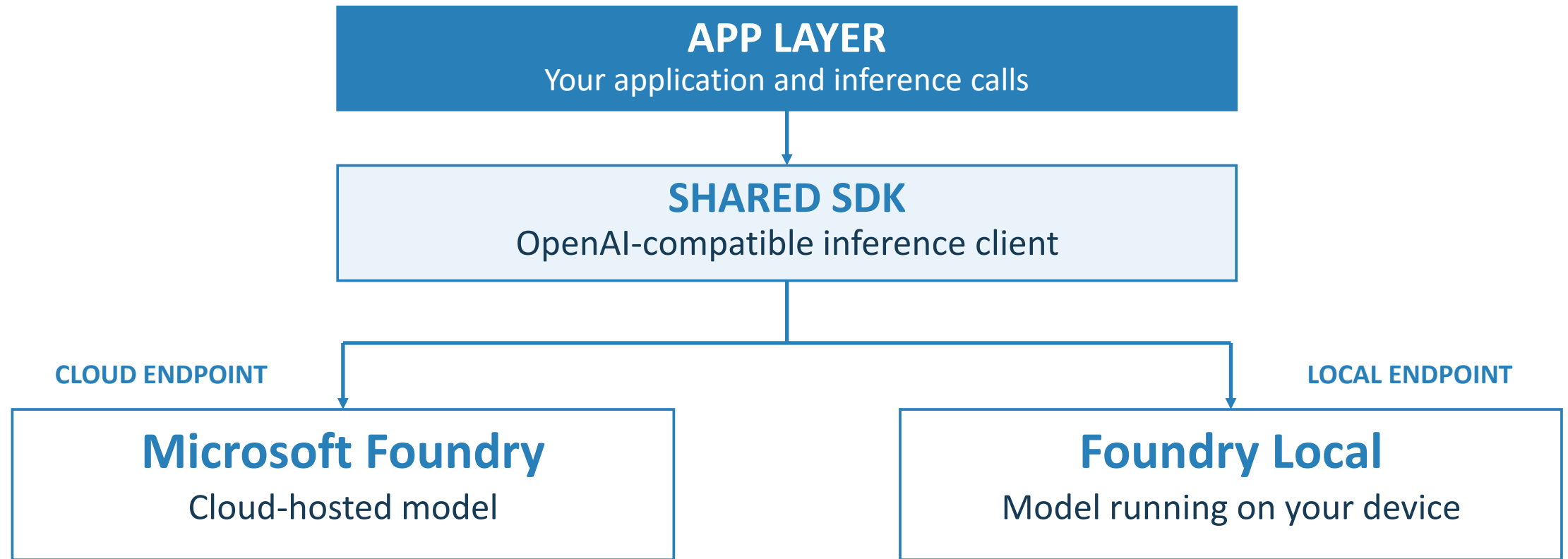
Massimo Bonanni

Senior Technical Trainer, Microsoft



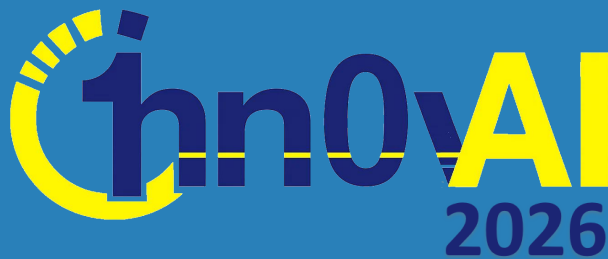


One app, cloud or local inference



Same application code. Different connection settings.

Configure endpoint, model and authentication. Use compatible inference APIs and model capabilities.



Microsoft Foundry

The End-to-End Platform for AI Applications



The Challenges of AI



Proliferation of Models

Different APIs from multiple providers without a unified access layer



Fragmented tools

6+ separate tools just to create an AI application in production



Security Risks

API keys scattered in env files, inconsistent authentication patterns



Evaluation

No standard way to measure quality, consistency, or safety



Poor compliance

No built-in guardrails for potentially harmful AI content



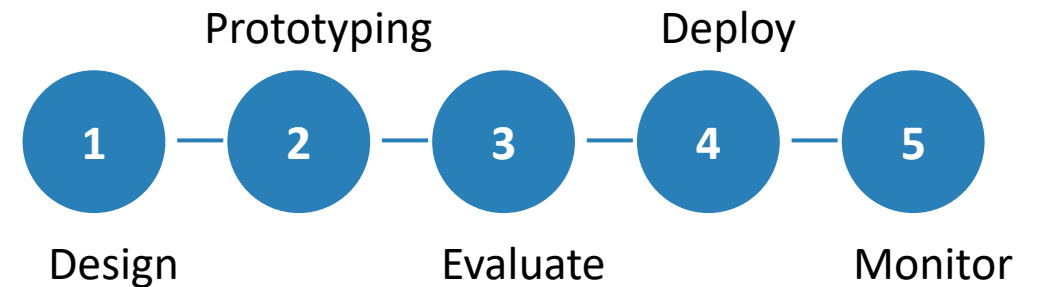
Opaque Costs

No unified view of token usage, model costs, and deployment expenses

What is Microsoft Foundry?

The unified, end-to-end platform for building, evaluating, and deploying AI applications at enterprise scale.

GenAIOps Lifecycle



10k+ Models

Access OpenAI, Meta Llama, Mistral, Phi, and more from a single workspace



Unified SDK

One SDK for inference, agents, evaluation, and content safety



Enterprise-Grade

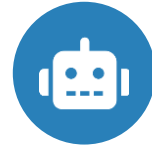
Managed Identity, Entra ID, RBAC and integrated guardrails for Responsible AI

Features at a glance



Model Catalog

10K+ models: OpenAI,
Llama, Mistral, Phi



AI Agents

Code Interpreter, MCP,
Bing, SharePoint



Rating

15+ built-in evaluators for
quality and safety



Guardrails

4 damage categories,
severity scale 0-7, jailbreak,
PII ...



Monitoring

OpenTelemetry, App
Insights, tracing



Security

Managed Identity, Entra ID,
RBAC



Integrations

AI Search, Cosmos DB,
Fabric, GitHub

AI Models Catalog



GPT-4o, GPT-4, GPT-3.5,
DALL-E, Embeddings



Call 3.1, Call 3
Multiple sizes



Mistral Large, Mistral
7B, Codestral



Phi-3, Phi-3.5
Optimized SLMs

...

Model Deployment Options

Serverless API

- For Foundry Models sold by Azure and select partner or community models.
- Pay-per-token (Standard) / Monthly fee (PTU)
- No infrastructure to manage
- Different processing scope (regional/Geo/Global)

Managed Compute (Preview)

- For open-source, partner, and custom models.
- Dedicated compute instances
- Performance guaranteed by SLA
- Preview

Deployment types: match economics to workload

Standard

PAY PER TOKEN

Best for Variable or unpredictable traffic

Behavior Development, testing, and production

Scope Global • Data Zone • Regional

Provisioned

RESERVED CAPACITY

Best for Sustained, high-volume demand

Behavior Predictable throughput and latency

Scope Global • Data Zone • Regional

Batch

DISCOUNTED ASYNC

Best for Large, non-urgent workloads

Behavior Bulk jobs without real-time latency

Scope Asynchronous processing

Developer

EVALUATION

Best for Fine-tuned model evaluation

Behavior Low-friction testing—not production

Scope Availability varies by model

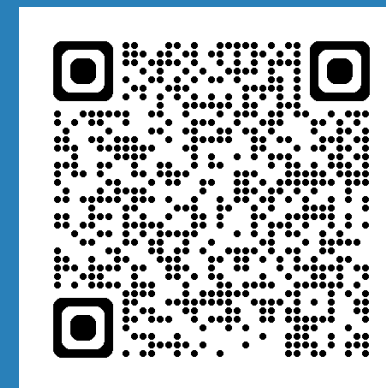
Data stored at rest remains in the designated Azure geography; inference processing location depends on deployment scope.



DEMO

Hybrid Ping Pong

The cloud side





Foundry Local

Hosting for local models



Executive summary

The product family is maturing along two distinct deployment paths.

On-device runtime

Embed private, offline AI in client applications with native SDKs, automatic hardware optimization, and an optional OpenAI-compatible local service.

Enterprise edge runtime

Run inference on Arc-enabled Kubernetes with declarative model lifecycle, CPU/GPU options, multi-node scale, and disconnected operations.

Latest developer shift

Foundry Local v2.0.1 moves in-process development to a unified Session API across languages, while retaining OpenAI-compatible REST contracts.

Latest platform shift

Azure Local extension 2607 adds local model evaluation, multi-GPU model parallelism, and safer automatic GPU inference tuning.

Bottom line: treat “Foundry Local” as a portfolio decision, not one runtime. Start with workload location and operating model.

Two deployment paths

Same goal: local inference. Different target, scale, and operational model.

Foundry Local

Client and application runtime

- Runs on end-user devices
- SDK embedded in the app
- Offline and no Azure subscription required
- Automatic CPU/GPU/NPU selection
- Best for desktop, mobile, and local tools

Foundry Local on Azure Local

Enterprise on-premises inference platform

- Runs on Arc-enabled Kubernetes
- Operator and custom resource lifecycle
- Connected or disconnected environments
- CPU/GPU, multi-node, high-parameter models
- Best for shared and regulated workloads

Choose by workload boundary

On-device runtime: what is new

Version 2.0.1 is a meaningful SDK and packaging transition.

1

Unified Session API

Model → Session → Request/Items → Response across C#, Python, JavaScript/TypeScript, Rust, C++, and C ABI.

2

Multimodal + streaming

Typed text, image, audio, tensor, tool-call, and tool-result items, plus sync/streaming and cancellation.

3

Tool calling

Register tools at session level and exchange typed tool-call and tool-result objects.

4

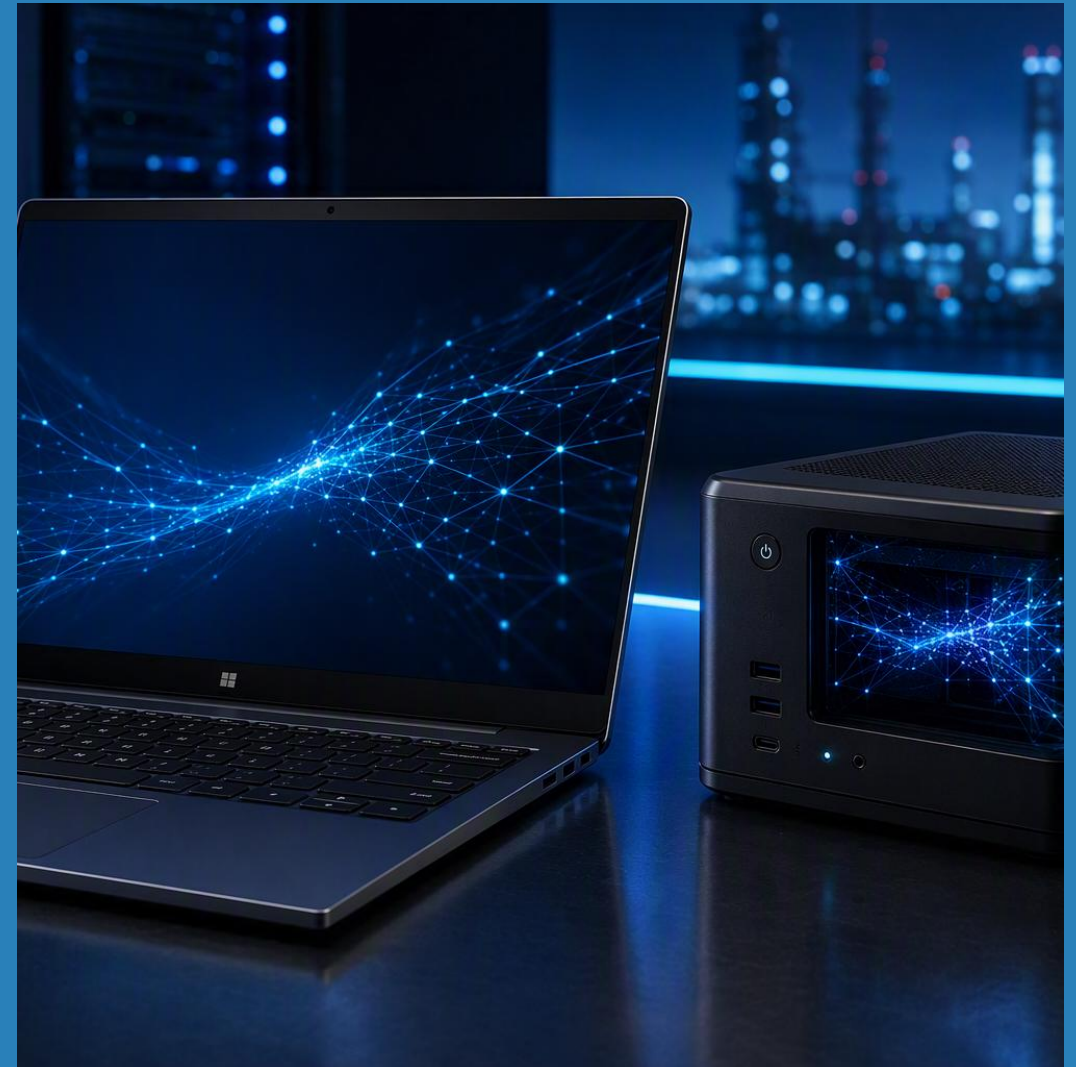
Simpler packaging

One package per SDK with automatic execution-provider selection; separate WinML packages are removed.

5

Migration action

New apps should use ChatSession, EmbeddingsSession, or AudioSession. Direct in-process OpenAI-style clients are deprecated.



Azure Local: latest release train

The 2605 → 2607 sequence shifts the platform from foundational scale to production-oriented operations.

2605 | June

- Multi-node Kubernetes deployment
- Disconnected operations
- vLLM runtime
- Model caching + automatic GPU tuning

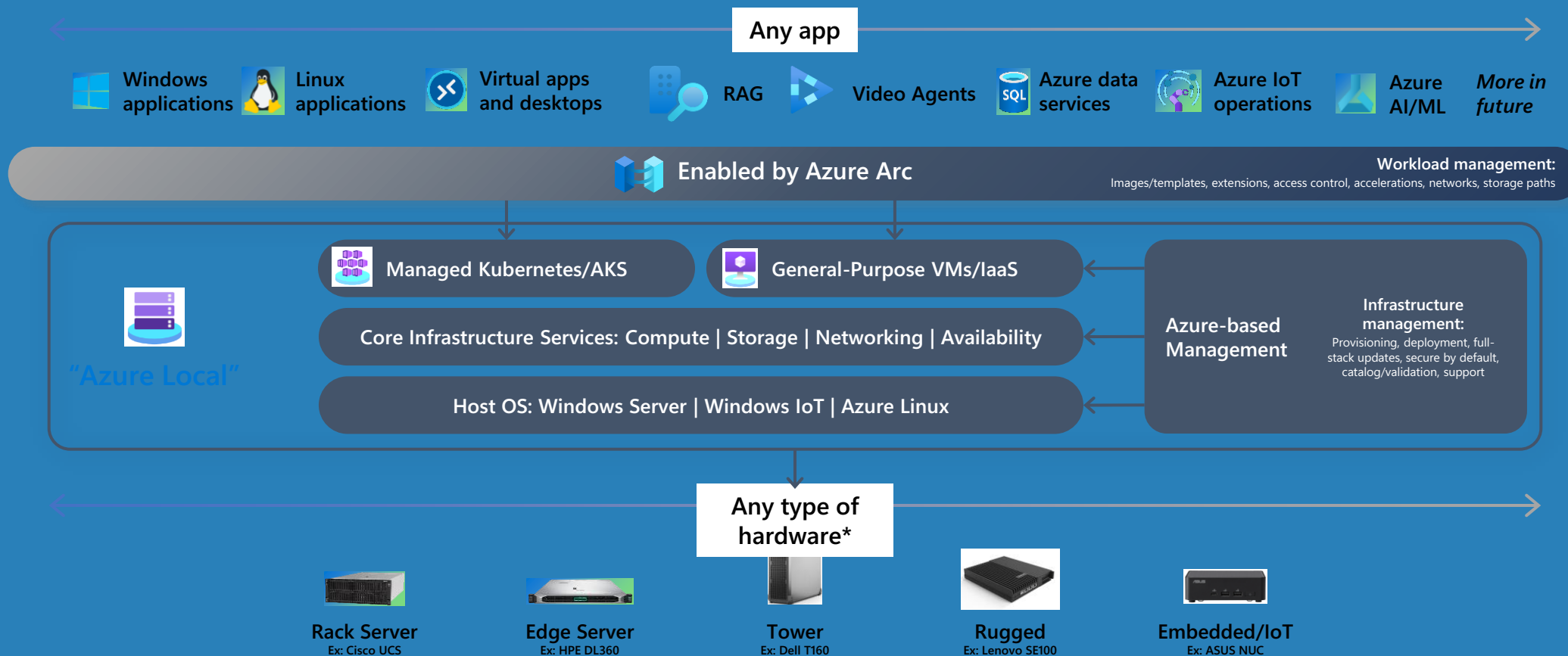
2606 | July

- Kubernetes Gateway API
- Inference-aware Endpoint Picker
- External TLS termination
- Disconnected networking bundle

2607 | August

- On-cluster model evaluation
- Tensor and pipeline parallelism
- Enhanced GPU memory profiling
- Auto-fit maximum context length

Validated on Azure Local, supporting other infrastructure



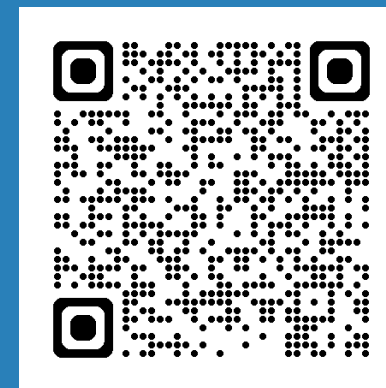
*Must meet minimum requirements per operating system and solution-level pass validation



DEMO

Hybrid Ping Pong

The local side



Microsoft Foundry vs. Foundry Local



Microsoft Foundry

Cloud • ai.azure.com



Foundry Local

On-Device • ~20 MB runtime

Models	10K+ (GPT-4o, Llama, Mistral...)	Hardware-specific SLMs (Phi, Qwen, DeepSeek)
Internet	Required	Not necessary
Latency	Network Round-trip	Zero network latency
Privacy	Data processed in Azure	Data doesn't leave the device
Cost	Pay-per-token / Reserved capacity	Free, no API key
Scale	Multi-user, enterprise	Single-user, on-device
Enterprise	RBAC, safety, monitoring, audit	None (lightweight runtime)
Fine-tuning	Supported	Not supported



Questions?



Marco Dal Pino

Senior Technical Consultant, Microsoft

info@contoso.blog

<https://www.linkedin.com/in/marcodalpino/>



Massimo Bonanni

Senior Technical Trainer, Microsoft

massimo.bonanni@microsoft.com

<https://info.massimobonanni.it>

