



AI INFERENCE ON AKAMAI CLOUD

Akamai AI Orchestrator

AI inference as a global compute cache

● TECHNOLOGY DEMO

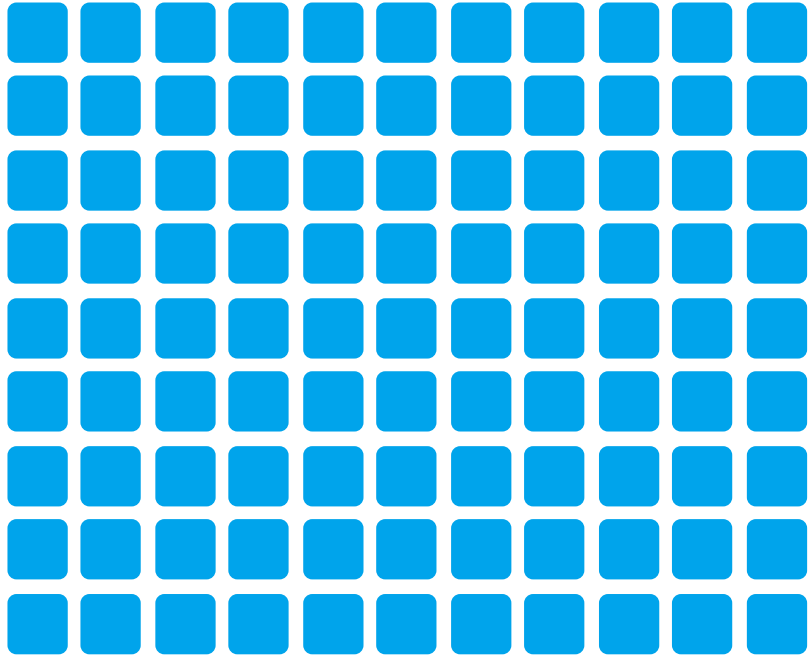
Du'An Lightfoot

Sr. Developer Advocate, Akamai

AGNTCon + MCPCon North America 2026

The problem

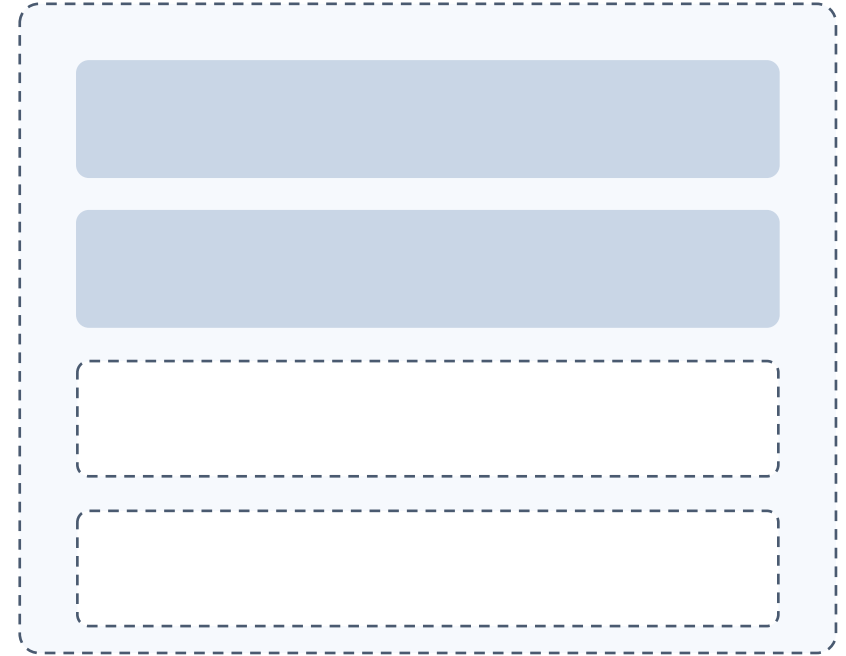
99 models



?



One GPU



Which models, in which city, at what time?

Treat AI inference like a global compute cache

Every request is scored on three things: is the model loaded, how close, how busy

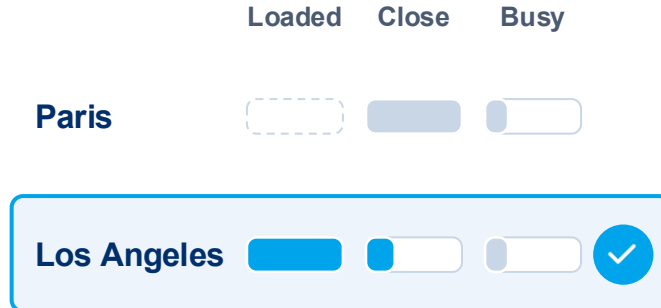
1 Cold start

Nothing is loaded, so the closest region wins. The first call pays the load.



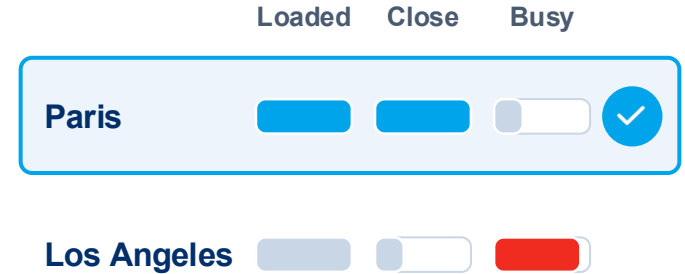
2 Affinity

A loaded model beats a closer region, even across an ocean.



3 Rebalancing

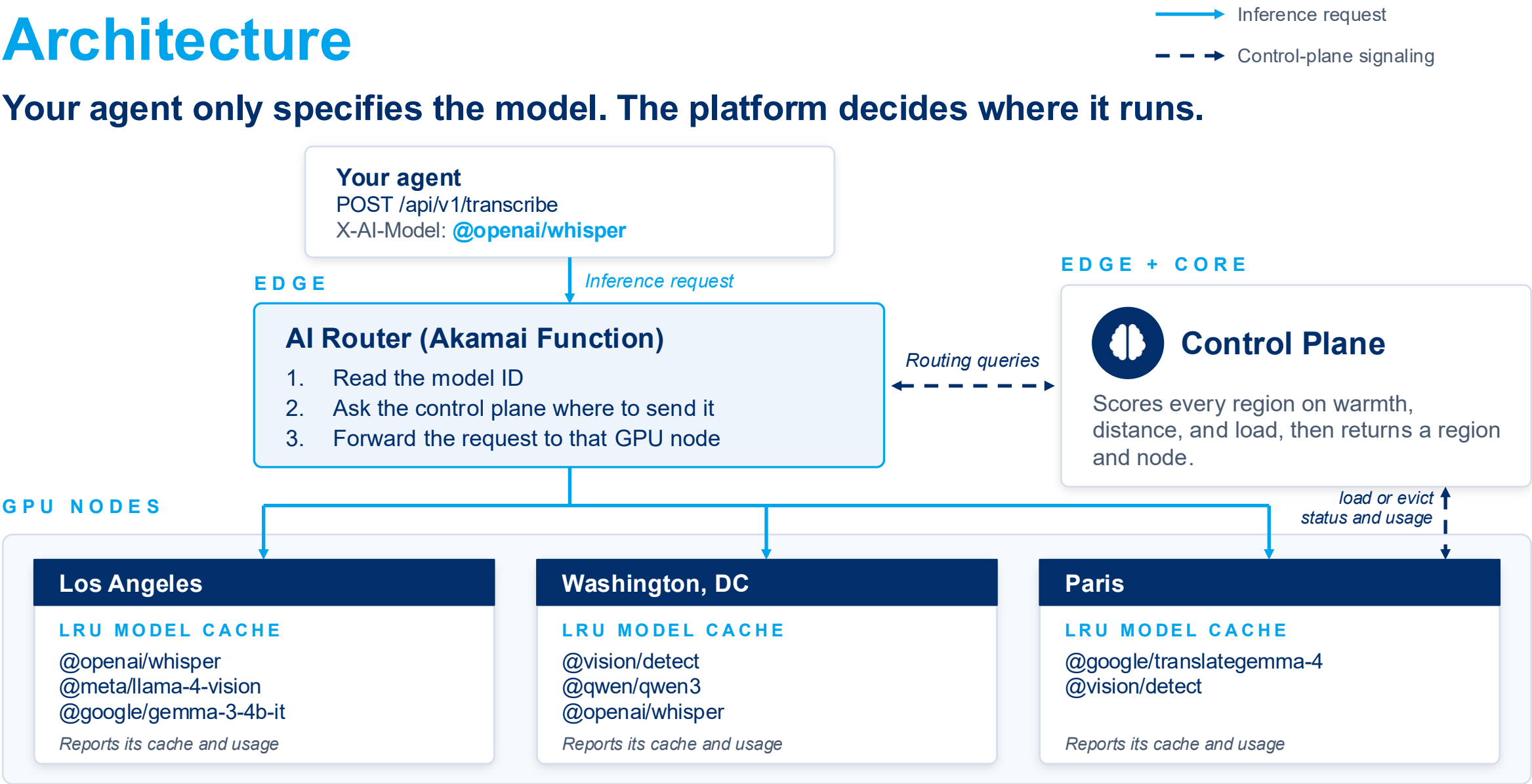
Demand grows in Europe. Paris loads the model and traffic comes back.



Models end up where the demand is.

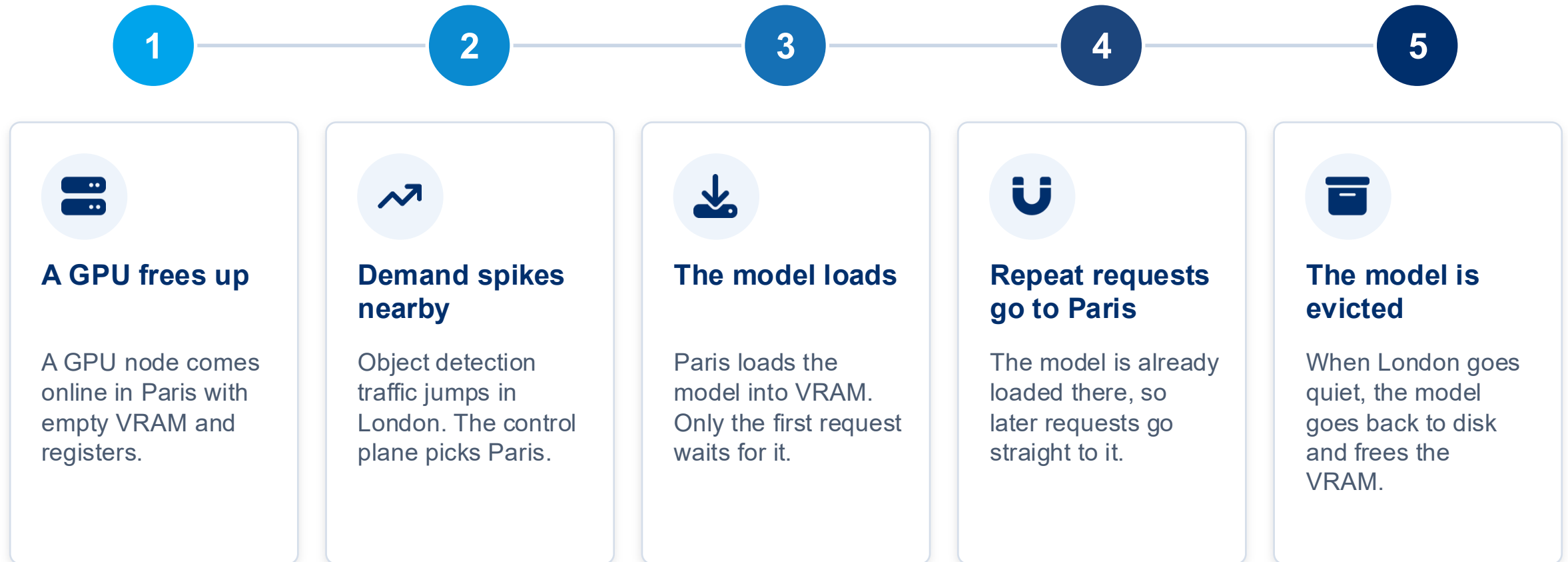
Architecture

Your agent only specifies the model. The platform decides where it runs.



Life of a request

How a model ends up close to the demand



ChromeFileEditViewHistoryBookmarksProfilesTabWindowHelp

Multi-Modal Inference — AI

cd ee0c83-6d9f-450e-ac35-3a46483ce02a.fwf.app/multimodal/

☆Incognito (3)

Akamai AI Orchestrator Examples

Multi-ModalChatEmbeddings

Multi-Modal Inference

Object detection and audio transcription running simultaneously through the AI router. Video frames are sent to YOLO, audio chunks to Whisper — each routed to the best available GPU node.

AI ROUTER ENDPOINT

YOLO MODEL

WHISPER MODEL

FRAME INTERVAL

https://a4bfe2b3-d2db-482f-bc65-8766a994807c.fwf.api

YOLOv8 Large

Large v3

2s

AUDIO CHUNK

5s

STATUS

Idle

DETECTIONS

0

TRANSCRIPTIONS

0

YOLO Latency

detect

ms

Whisper Latency

transcribe

ms

MIN — AVG — MAX — P95 —

SAMPLES 0

MIN — AVG — MAX — P95 —

SAMPLES 0

Start

Stop

Requires camera & microphone access.

Object Detection

YOLO

Akamai AI Orchestrator

DashboardModelsHow It WorksReset3 regionsConnected

+ -

ROUTING WEIGHTS

Distance3.0

Load1.0

Warmth1.0

Datacenters

Hide Nodes

fr-par-2

CUDA21d 8h

48.86, 2.35

UTILIZATION

0.0%

0 active · 0 pending

MEMORY

7.5%

12.9 GB / 172.9 GB

VRAM

1.6%

1.6 GB / 95.6 GB

GPU

0%

NVIDIA RTX PRO 6000

Blackwell Server Edition · 33°C

ca-central-hybrid

CPU114d 12h

43.71, -79.41

UTILIZATION

0.0%

0 active · 0 pending

MEMORY

9.9%

1.5 GB / 15.6 GB

COMPUTE

cpu

CPU inference

Muon Ops

MODELS · CACHE 0.0% (0 B / 12.5 GB)

● warm ● cold ● loading

What this means for your own agents

1 Put the model name in the request and nothing else.

No region, no endpoint, no GPU. If your agent hardcodes where the model lives, nothing can move it.

2 Loading a model is the slow part, not the network.

A model that's already loaded 5,000 miles away answers faster than one that has to load next door.

3 If an agent can change your routing, decide what it's allowed to touch.

Some changes it can make on its own. Some need a person. Write that down before you ship it.

GET STARTED

Build on Akamai Cloud

Scan to create an account and get the code samples.



Create an Akamai Cloud account

Includes \$300 in credit

login.linode.com/signup



Get the code samples

Akamai Developers on GitHub

github.com/akamai-developers

RESOURCES

- ✓ **\$300 cloud credit**
Free for new accounts
- ✓ **Akamai Developer**
akamai.com/developer
- ✓ **Docs and guides**
techdocs.akamai.com
- ✓ **Code samples**
github.com/akamai-developers

Come find us at the Akamai booth

