

UNITEDGE® AI • PRIVATE AI APPLIANCE

Your AI factory. In 3U.

Frontier AI, virtual machines, containers and serverless in one appliance at your site. Your data stays put. Your cost stays fixed.



1 TB

AI memory for frontier and open models

0 bytes

Of prompt data sent off site

2 sides

Active-active, no single point of failure

1 visit

To go live at your first site

AI is moving to where your data lives.

The most valuable AI work happens next to the most sensitive data: in clinics, plants, trading floors and stores. That is exactly where public AI services cannot follow.

01

Data can't leave

Patient records, plant telemetry, trading data and IP are bound by regulation and contract to stay on site. Public AI APIs are off the table.

02

API bills never stop

Per-token pricing grows with every new use case. The more value AI creates, the more it costs, and nothing is ever owned.

03

No one on site to run it

Branches, plants and clinics have no AI or infrastructure team. Whatever goes there must operate unattended and be fixed remotely.

The answer: a private AI factory small enough for the edge, and autonomous enough to run itself.

Three clouds in one box.

Two AI Engines and two Compute Engines, side by side in one 19-inch chassis. Everything is doubled, so nothing has to stop.

1 TB AI memory	64 Compute Engine cores	1 TB ECC RAM	2×3.84 TB Clustered NVMe
4×25 GbE LAN and WAN	2 feeds ~1.4 kW peak	3U 19-inch, 23 in deep	Zero-touch Enrolls on first boot

INCLUDED

High-performance inference

Frontier and open models on 1 TB of AI memory, served as private, OpenAI-compatible endpoints.

Frontier LLMs Vision SLMs Private API

INCLUDED

Cloud native at the edge

Run apps where the data is: VMs, Kubernetes and serverless, highly available across both Compute Engines.

VMs K8s pods Serverless HA

INCLUDED

Software-defined networking

Virtual firewalls, load balancers and routers, configurable per tenant and highly available.

Firewall Load balancer Router SDN

Run the leading open models on site.

A curated catalog of open-weight models, security-scanned and license-reviewed, deployed to any appliance in one click and served behind one private, OpenAI-compatible gateway.

Model catalog
Open-weight models curated, security-scanned and licence-reviewed by UnitedLayer - deploy to any appliance in one click

Search sites, workloads, models... Meridian Health - Edge JR

Upload fine-tune Request a model

Origin: US EU China

POLICY Tenant Meridian Health: China-origin models allowed in Research workspace only - all models run on-site, no external calls

Model	Provider	Size	Deployment Status
GLM-5.2	Z.ai - Frontier	744B MoE	Deployed - 9 sites
DeepSeek V4 Pro	DeepSeek - Frontier	1.6T MoE	Available
DeepSeek V4 Flash	DeepSeek - Frontier	284B MoE	Deployed - 12 sites
Kimi K2.5	Moonshot AI - Frontier	1T MoE	Available
gpt-oss-120b	OpenAI - Frontier	117B MoE	Deployed - 6 sites
Llama 4 Maverick	Meta - Multimodal	400B MoE	Available
Nemotron 3 Super	NVIDIA - Enterprise	120B MoE	Deployed - 3 sites
Mistral Large 3	Mistral AI - Frontier	675B MoE	Deployed - 2 sites
Qwen 3.5 397B	Alibaba - Frontier	397B MoE	Available
Qwen3-Coder 480B	Alibaba - Coding	480B MoE	Available
Devstral 2	Mistral AI - Coding	123B	Deployed - 4 sites
Gemma 4 31B	Google - Vision	31B	Deployed - 7 sites

The model catalog in the UPC Vektor console: filter by origin, deploy in one click.

US, EU and China

Frontier, vision, coding, speech, industry and small models.

Policy per tenant

Allow or block by origin, license or workspace.

Change one URL

Keep your code. Fail over to the peer AI Engine, then your core AI Factory.

AI endpoints
Private, OpenAI-compatible endpoints on your edge appliances - one key, one gateway, every site

Search sites, workloads, models... Meridian Health - Edge JR

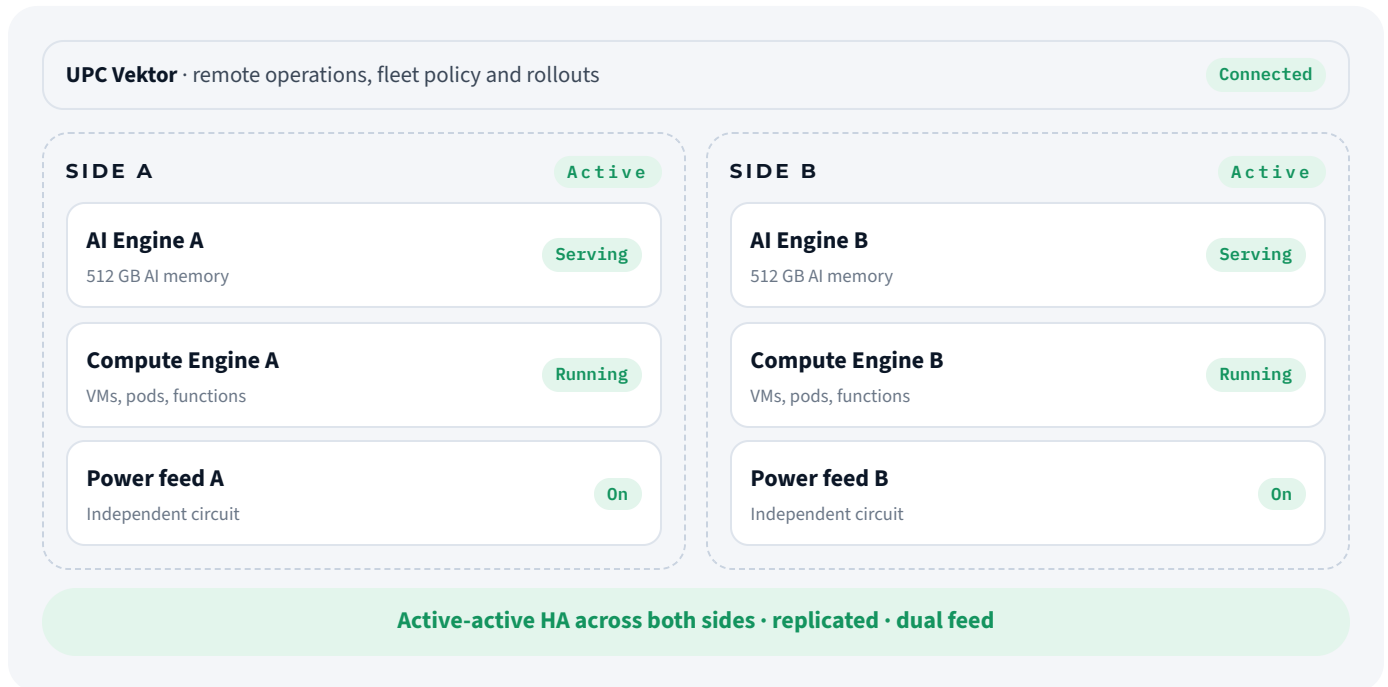
+ New endpoint

Endpoint	Model	Site	Tokens / s	PPS First Token	Requests 24 h	Failover Path	Status
clinical-notes-assist	GLM-5.2	SFO-Branch-01	1,240	410 ms	18.4 K	Local → peer → AI Factory	Live
support-copilot	DeepSeek V4 Flash	SFO-Branch-01	2,860	220 ms	52.1 K	Local → peer	Live
plant-vision-qa	V4 Flash + vision	DEN-Plant-02	940	260 ms	9.7 K	Local → AI Factory	Live
contracts-review	DeepSeek V4 Pro	NYC-Trading-03	610	690 ms	3.2 K	Dual-engine	Live
call-transcribe	Whisper large-v3	ATL-Hub-01	-	120 ms	27.5 K	Local	Live

Licenses vary by model.

Built to run with no one on site.

Active-active across two sides of one appliance. Lose a feed, an engine or a cable, and service continues.



No single point of failure

Lose a feed, an engine or a cable; service continues.

Failover in seconds

Firewall, load balancer and AI endpoints switch automatically.

Ruggedized to ship

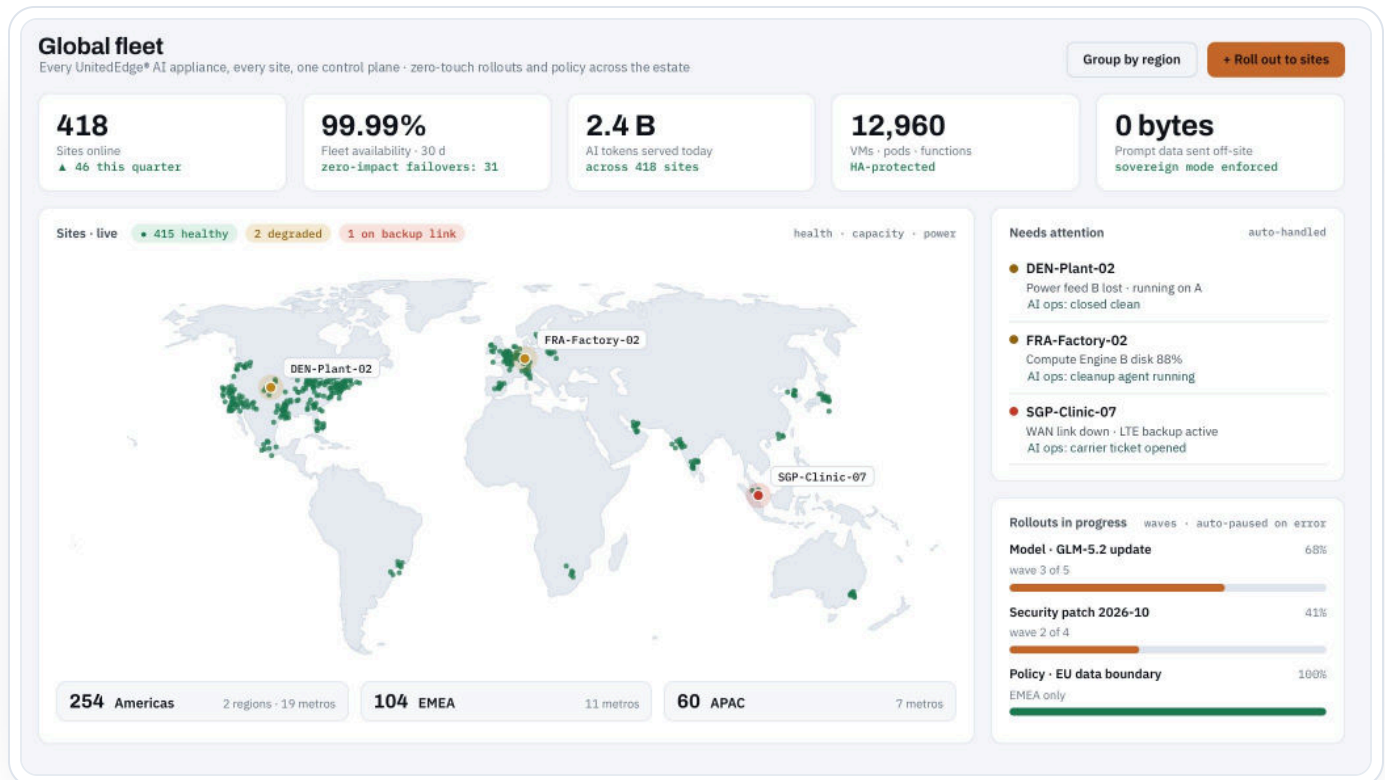
Every connector locked, every mass shock-mounted.

Lights-out operation

Remote console, power control and LTE backup.

Every appliance, one fleet, one console.

UPC Vektor manages one site or hundreds: self-service for VMs, Kubernetes, serverless and AI endpoints, with fleet policy, rollouts in waves, access and audit.



- | | | | | |
|---|---|---|---|---|
| <p>01</p> <p>Enroll</p> <p>Zero-touch, pulls its policies</p> | <p>02</p> <p>Configure</p> <p>By region, site or tenant</p> | <p>03</p> <p>Update</p> <p>Waves, auto-pause on error</p> | <p>04</p> <p>Observe</p> <p>Health, capacity, power, carbon</p> | <p>05</p> <p>Remediate</p> <p>AI ops auto-fix, remote console</p> |
|---|---|---|---|---|

Autonomous ops for super-productive IT

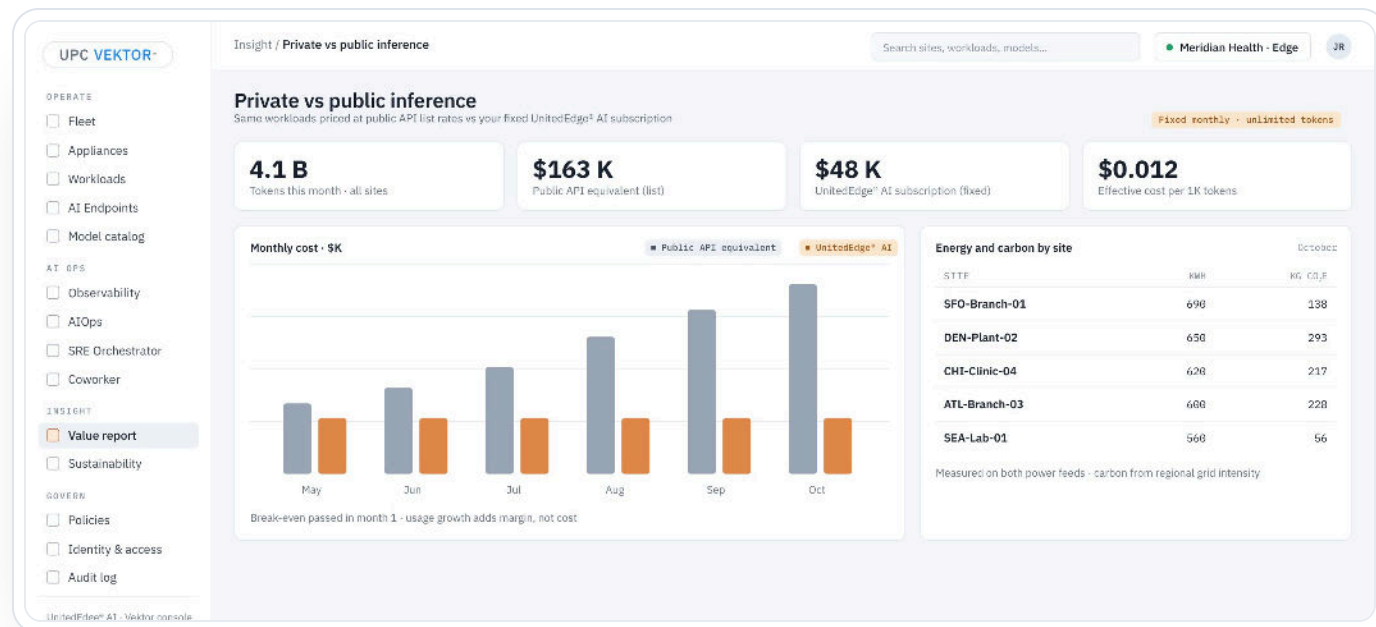
The SRE Orchestrator finds and fixes issues at every site, with people approving by policy. AI Coworker answers in plain language, acts and updates tickets.

- | | | | |
|----------------------------------|---------------------------------|------------------------------------|---------------------------------|
| 85%
L1/L2 closed by AI | 78%
Faster to resolve | 3×
Engineer productivity | 70%
Staff hours saved |
|----------------------------------|---------------------------------|------------------------------------|---------------------------------|

Outcomes reported in customer deployments of UnitedLayer's autonomous operations platform.

One fixed price. Unmetered tokens.

Appliance and platform in one subscription, with managed operations optional. As AI use grows across your sites, your cost stays the same.



Value report: the same workloads priced at public API list rates versus the fixed subscription. Sample figures, not a quote.

Sovereign by construction, not by policy.

Data stays on site

Models, prompts, outputs and embeddings never leave the appliance.

Isolated tenants

Compute, storage, network and AI per business unit.

Encrypted at rest

Hardware-sealed keys, secure boot and signed updates.

Tamper-aware

Lid, bezel and shock sensors raise alerts.

Zero-trust access

Your identity provider, role-based access and full audit.

Recoverable

Immutable cyber vault backups, DR in UnitedLayer facilities.

Live at your first site in a single visit.

STEP 1

Size and order

Use cases, models and workloads agreed in a workshop.

STEP 2

Ships ready

Pre-imaged, burned in and tested, in a shock-rated crate.

STEP 3

Plug in

Two power cords and two network handoffs per side.

STEP 4

Enrolls itself

Joins the console, pulls your policies and starts serving.



Edge plus core

Appliances at your sites and the UPC AI Factory in your core, under one console and one set of policies.

Next steps

- 1 **Workshop:** pick two or three AI workloads and the VMs to consolidate.
- 2 **Pilot:** one appliance at one site, run from the console.
- 3 **Scale:** across sites under one subscription.

**Pick two workloads.
We'll bring the box.**

CALL

+1-888-853-7733

EMAIL

contact@unitedlayer.com