Infra

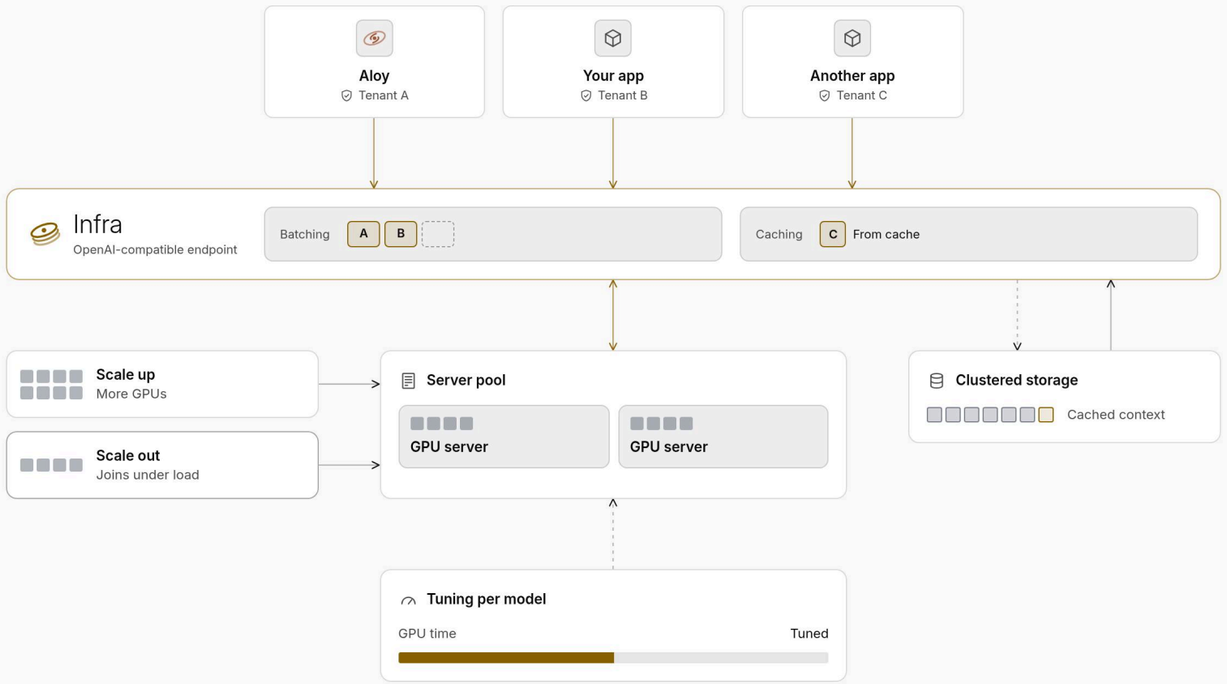
Infra is the inference layer that runs your AI models. It scales from a single GPU server to a multi-node cluster while exposing one consistent OpenAI-compatible API to your applications.

This overview follows the page at rhenium.tech/products/infra/, as of November 20, 2026.

One endpoint for every model

Applications connect to Infra through a single OpenAI-compatible endpoint. Both Aloy and your own internal applications call the same interface.

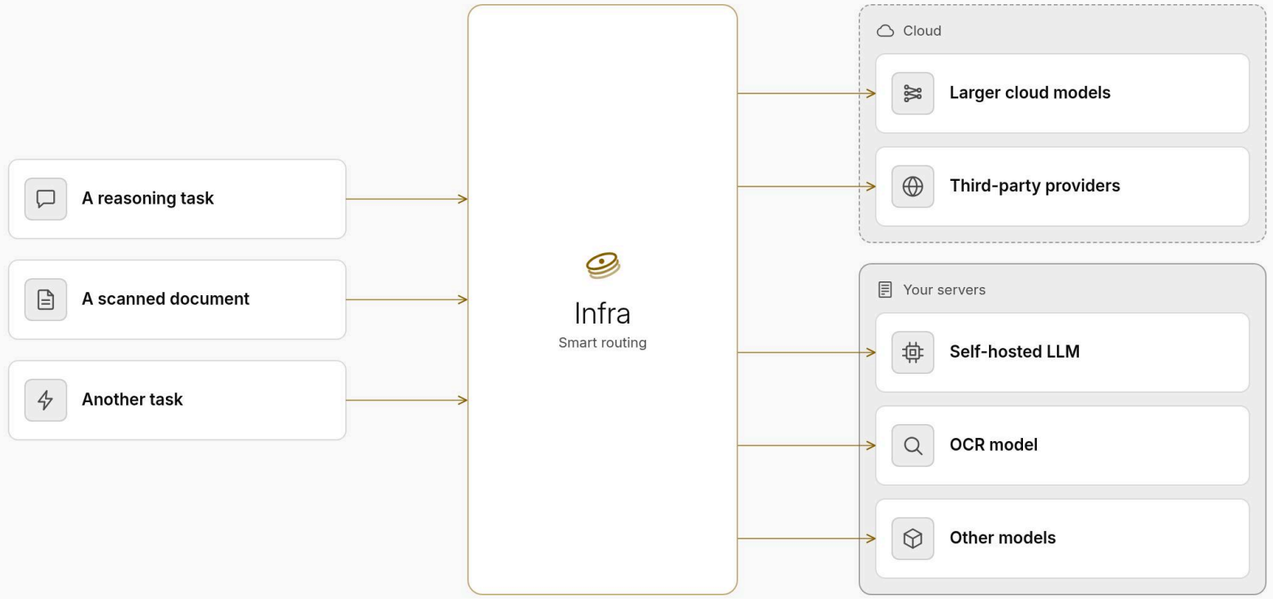
Behind that endpoint, Infra pools GPU capacity across servers and isolates workloads per tenant. Applications do not need to change when underlying hardware grows or shifts.



Routing across local and cloud models

Different workloads require different models: reasoning tasks need larger language models, while document extraction calls for dedicated OCR or smaller specialized models. Infra routes each request to the appropriate model automatically.

Models can run on your own servers, in the cloud or through third-party providers. Applications still interact with one unified endpoint under one policy framework.



Built for low latency and lower GPU cost

Inference cost scales directly with GPU time. Rhenium profiles and optimizes each deployed model for your target hardware to reduce latency and increase throughput.

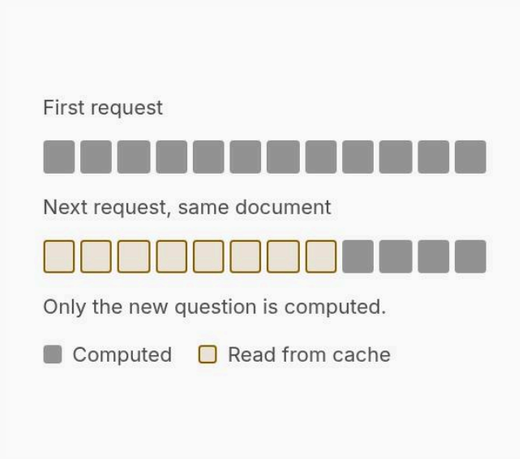
Infra also handles distributed caching, request batching and model optimization across nodes. Repeated context is reused rather than recomputed, reducing GPU utilization, network load and operating cost.

Model-specific tuning



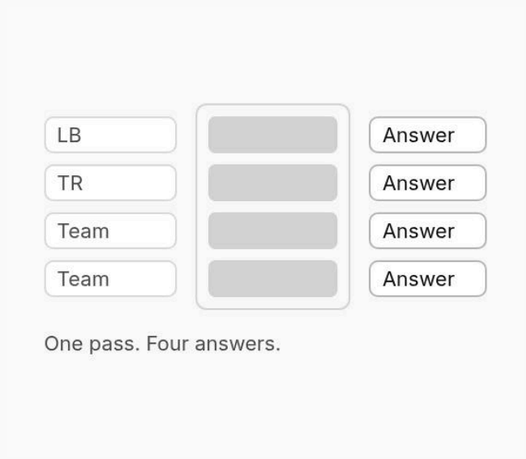
Each deployed model is profiled and tuned for its architecture and target hardware, cutting response latency without changing output quality.

Distributed caching



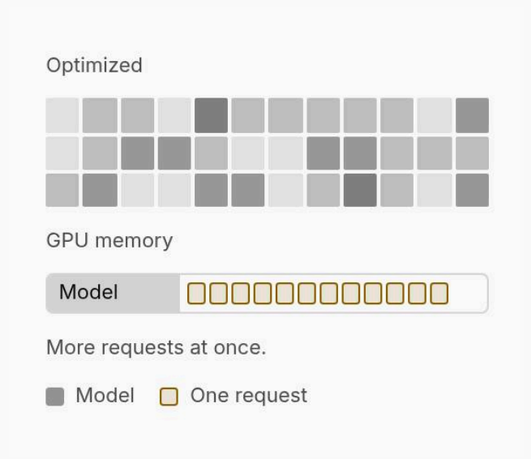
Shared context and repeated prompts are cached across inference nodes and storage so identical work is never recomputed.

Dynamic batching



Concurrent requests from across the organization are grouped and processed efficiently, maintaining steady throughput during peak demand.

Memory optimization

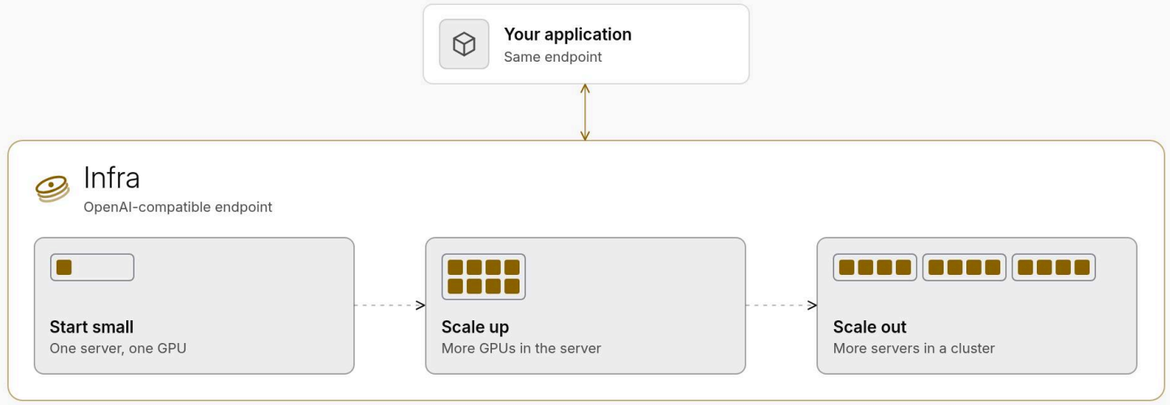


Models are calibrated to reduce memory footprint per request, allowing higher concurrency or smaller hardware footprints with verified accuracy.

Scaling from one GPU to a cluster

Deploy Infra on a single server with one GPU to start. Add GPUs to scale up within the host, or add servers to scale out into a cluster as adoption grows.

The same software stack runs from a single server to a data-center cluster, and your applications keep the same endpoint throughout.



Deployment modes

Deploy Infra paired with Aloy or standalone as your organization's inference layer.

Setup	What you get
With Aloy	Enforces Aloy's permission and tenant isolation model end to end.
Standalone	Provides a native OpenAI-compatible endpoint for your internal applications.
For system integrators	Embeds into client architectures from single-server deployments to full clusters.
Offline	Paired with Aloy and an offline policy, runs fully air-gapped with no external network traffic.

Talk to us

Every project starts with a conversation. Tell us what you are working on and where it gets stuck, and we will write back by email.

Email info@rhenium.tech

Web rhenium.tech/products/infra/