

Own your Inference

Five reasons to own your inference instead of renting it.
Rented tokens are a bill that never stops. Owned tokens get cheaper every quarter the server runs.

0 fees

per-token fees on owned capacity

4.32 TB

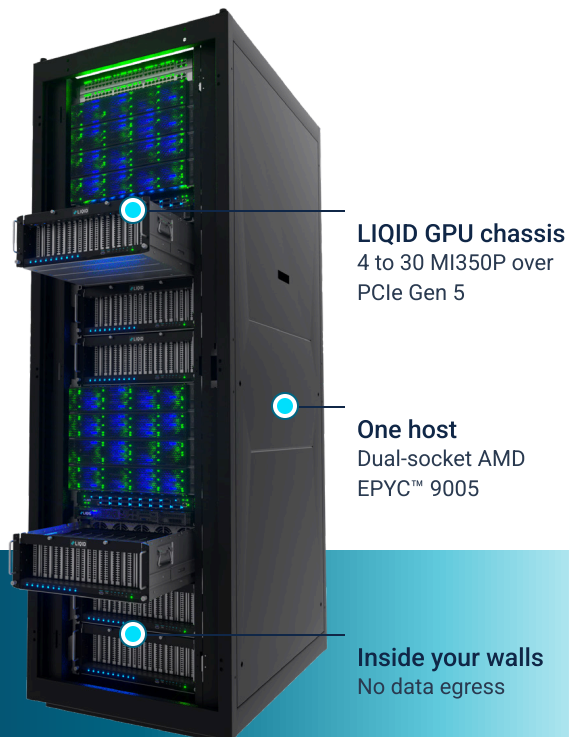
HBM3E on one server

30 models

in parallel, one per GPU

~22 kW

for 30 GPUs, air-cooled



Five Reasons to Own Your Inference

Here is when owning is worth it, and why it is easier than it used to be.

1

Your data never crosses a boundary you don't control

For regulated, classified, sovereign and clinical workloads this is a condition of operating. On-premises deployment removes the data-residency question from the architecture.

[AMD SEV trusted execution with EPYC™ 9005 · secure boot · DICE attestation](#)

2

Cost per token falls every quarter the server runs

A rented token costs what it costs, priced to someone else's margin. An owned server is an asset whose output grows while it depreciates, so unit cost falls as you use it more.

[No per-token fees on owned capacity](#)

3

Capacity is yours at peak, not rationed

Hosted capacity tightens exactly when demand spikes. Capacity you own is there when traffic arrives, and every model has its own GPU, so one busy job never slows another.

[Available now, from 4 to 30 GPUs](#)

4

It's predictable enough for a CFO

Metered consumption is the hardest AI line to forecast. A known capital cost against known power and known throughput is a number finance can model and depreciate.

[Capex you can schedule, not usage you have to explain](#)

5

The server fits the facility you already have

MI350P is an air-cooled PCIe card at 600 W, and a full 30-GPU system draws about 22 kW. No liquid cooling, no rack redesign, no new power distribution.

[One host instead of a multi-server build-out](#)

Tell us what you spend on hosted inference today and which workloads can't leave your estate. We size the equivalent on-premises system, power and footprint.

Built for

Regulated enterprises · Public sector · Healthcare · Financial services · Sovereign AI programs

Proven on

AMD Instinct™ MI350P · AMD EPYC™ 9005 · ROCm™ · 200+ Liqid deployments

Model Your Costs

sales@liqid.com · +1 303.500.1551