



Every revenue cycle agent. PHI inside.

Every agent. One server.

A revenue cycle agent suite on AMD Instinct™ MI350P. Prior auth, denial appeals, eligibility and documentation on one Liquid UltraStack server, with PHI inside your walls.

20 GPUs

MI350P on one host, room for 30

2.88 TB

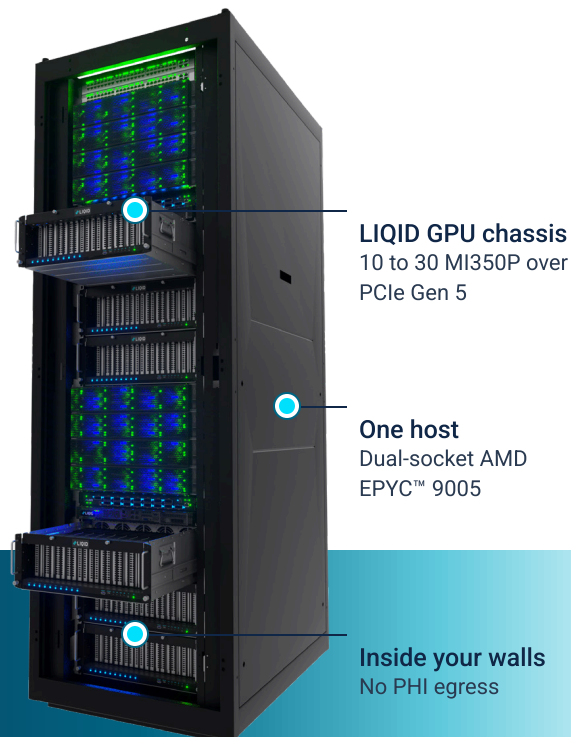
HBM3E on one server

4 agents

plus 4 flex GPUs on one server

0 hops

network hops between GPUs



The Revenue Cycle Gap

Clinical AI is mainstream. Revenue cycle AI is still a stack of competing pilots.

Adoption Gap

Clinical AI is everywhere

75% of health systems run AI; ambient notes went from 42% to 68% in a year.

Revenue cycle lags

Only 28% use prior auth AI and 25% denial prediction.

Pilot Sprawl

Every agent is its own pilot

Each one competes for its own GPUs.

The ROI is already proven

59–71% of adopters report 2× or better ROI.

PHI Pressure

PHI can't go to a cloud API

Not at the scale revenue cycle runs.

Staff own the submission

Every draft must cite the chart and the payer rule.

One Pool. Four Agents.

One 20-GPU Liquid UltraStack server runs four revenue cycle agents, each on its own GPUs, plus four flex GPUs for whichever queue surges. Every draft cites the chart and payer rule.

- » Fixed capacity replaces per-token fees
- » A CFO budget: denial rates, days in A/R and vendor fees
- » Grow from 10 to 20 to 30 GPUs on the same host

Pinned models · four agents, one pool

4
Prior auth

4
Appeals

4
Elig. + CDI

4
Docs

4
Flex

LIQID GPU chassis · 20 AMD Instinct™ MI350P · 2.88 TB HBM3E

PCIe Gen 5 connection · chassis to host

One dual-socket AMD EPYC™ 9005 host · one OS image

Prior Auth + Appeals · 8 GPUs

gpt-oss-120b drafts prior auths; Llama 3.3 70B drafts cited appeals.

1,256

tok/s, 72B-class, 8 GPUs

1 GPU

per 70B model at FP8

Eligibility + CDI · 4 GPUs

Mistral Small 3.2 and Llama 3.1 8B run eligibility and CDI checks.

9.5×

scaling, 1 to 10 GPUs

44,383

tok/s, 7B-class, 10 GPUs

Documentation + Flex · 8 GPUs

gpt-oss-120b summarizes notes; flex GPUs follow month-end surges.

~5B

active params, gpt-oss-120b

~65 GB

at MXFP4, fits one card

Measured results: Liquid Phase-3 testing on 1 to 10 MI350P GPUs on one host. Model specs are published. Notes on page 2.

Built for

Revenue cycle · Health information management · CDI · Patient access

Proven on

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

Size Your Revenue Cycle

sales@liquid.com · +1 303.500.1551

Proof: Measured on MI350P

Four agents don't need four pilots: **one pool of 20 GPUs serves them all, and flex capacity follows the queue.**

Result	What was measured	Why it matters here
1,256 tok/s on 8 GPUs	72B-class model, FP8, under interactive limits: P99 TTFT ≤ 2 s, ≥ 30 tok/s per user	Appeal letters and prior auth drafts at interactive speed for staff
9.5× scaling 1 to 10 GPUs, 7B-class	4,653 tok/s on 1 GPU to 44,383 tok/s on 10, one replica per GPU	Eligibility and CDI checks scale with every GPU you add
~5B active of 117B, gpt-oss-120b	Published model spec: only ~5B parameters active per token, ~65 GB at MXFP4	Reasoning-class model on one card, expected to decode at least as fast as the 72B result
What it means	Four agents, one pool	No agent waits for another pilot's GPUs

7B-class: FP4, one replica per GPU, concurrency 128. 72B-class: FP8, one replica per GPU, P99 time-to-first-token ≤ 2 s, ≥ 30 tok/s per user. Different test conditions; don't compare them. Source: Liquid Phase-3 testing. gpt-oss-120b, Mistral Small 3.2 and Llama performance is projected from these results; 20-GPU performance is not yet published.

Grow on the Same Server: 10, 20 and 30 GPUs

Add GPUs and a new job, not a new cluster.

10 GPUs 1.44 TB HBM3E Health-system-wide clinical documentation Scale summarization from a pilot unit to every clinician, sized for shift-change peaks. MODEL LAYOUT 1 × speech-to-text (Whisper-class, US) 7 × gpt-oss-120b note drafting 2 × coding suggestions and failover WHO BUYS, WHY NOW CMIO / CNIO. Ambient pilot succeeded; PHI can't go to a cloud API at scale.	20 GPUs 2.88 TB HBM3E Revenue cycle agent suite Prior auth, denial appeals, eligibility and CDI as separate agents sharing one pool. MODEL LAYOUT 4 × prior auth (gpt-oss-120b) 4 × appeals (Llama 3.3 70B) 4 × eligibility and CDI (Mistral Small 3.2, Llama 3.1 8B) 4 × documentation · 4 × flex WHO BUYS, WHY NOW CFO / VP Revenue Cycle. Denial rates, payer rule changes, RCM vendor fees.	30 GPUs 4.32 TB HBM3E Research AI and HPC on one server Literature and trial-matching agents on a frontier model, with GPUs shifted to FP64 simulation as needed. MODEL LAYOUT 8 × Mistral Large 3 (675B, FP8, multimodal) 12 × trial matching and literature workers 10 × shared between AI and FP64 WHO BUYS, WHY NOW CIO, academic medical center / VP R&D Informatics. Research computing consolidation; data can't leave.
---	---	---

MI350P runs FP64 as well as FP4, so at 30 GPUs one server can serve research computing and AI. Starting with one job? See the 4-GPU Clinical Summarization and Coding brief.

What UltraStack Adds

- » **One model per GPU:** 144 GB HBM3E holds a full 70B model at FP8, or gpt-oss-120b.
- » **No network between GPUs:** peer-to-peer over PCIe Gen 5 to one EPYC host.
- » **Air-cooled:** about 22 kW for the full 30-GPU system.

Start Fixed, Add Flexibility

- » **One host:** one OS image, one license footprint, one thing to patch.
- » **Room to grow:** 10 to 20 to 30 GPUs on the same host.

What Is Proven Today

- » P2P bandwidth saturates the link; RCCL collectives pass, 1 to 10 GPUs.
- » ROCm integration complete, developed jointly with AMD.
- » 200+ production deployments, 85+ patents.

The Efficiency Case Targets, not measurements*	Up to 75% fewer servers	Up to 2.1× more tokens per dollar	Up to 1.8× better tokens per watt	Up to 50% lower power
--	-----------------------------------	---	---	---------------------------------

Two questions settle the size of the opportunity: **how many prior auths and denials does your team work each month, and what does each one cost?**

Bring Us Three Numbers

- 1 **Volume:** prior auths, denials and claims per month.
- 2 **Case size:** notes, labs and payer policies per case.
- 3 **Concurrency:** sessions and staff at peak, incl. month-end.

Request a Sizing Session

We size it in one working session.
sales@liquid.com · liquid.com

*Efficiency targets compare against multi-server deployments of the same GPUs; they are Liquid internal projections using AMD published specifications, and results vary by configuration and workload. Performance figures for the AMD Instinct MI350P are based on AMD published specifications and Liquid Phase-3 testing from 1 to 10 GPUs, and are subject to change. Performance at 20 and 30 GPUs, and for 400B+ models, is not yet published; 20- and 30-GPU layouts and frontier-model fit are sizing arithmetic on published model sizes, not benchmark results. Models shown are US and EU open-weight models. Not a medical device; outputs require clinician, coder or revenue cycle staff review, and payer submissions remain the customer's decision. Adoption and ROI data: Eliciting Insights, Feb 2026 (n=120).