

# Investigate every alert.

**A financial crimes agent on AMD Instinct™ MI350P.** Triage, evidence, narrative and QA for every alert on one Liqid UltraStack server, with SAR data inside your walls.

**10 GPUs**

MI350P on one host, room for 30

**1.44 TB**

HBM3E on one server

**9.5x**

triage scaling, 1 to 10 GPUs

**0 hops**

network hops between GPUs



**LIQID GPU chassis**  
10 to 30 MI350P over PCIe Gen 5

**One host**  
Dual-socket AMD EPYC™ 9005

**Inside your controls**  
No data egress

## The Alert Backlog Dilemma

Route alerts into a growing queue, or investigate every one on infrastructure you control.



### Volume Pressure

#### Alerts outpace hiring

Alert volumes grow faster than investigator teams can hire.

#### Evidence eats the day

Investigators spend most of their time gathering evidence across systems.



### Cost Pressure

#### \$35–40B a year on AML

What U.S. financial institutions spend on AML operations.

#### Agents multiply tokens

Agentic workflows use 10–100x more tokens per task than one AI call.



### Data Pressure

#### SAR data can't leave

Customer and SAR data can't go to a hosted model.

#### Models must be validated

Pinned, self-hosted models are easier to take through model risk review.

## Triage It. Draft It. Sign It.

One 10-GPU Liqid UltraStack server runs every step after an alert lands. A small model routes it, a 70B model drafts the cited narrative, and a reasoning model checks it against known typologies.

- » Investigators review, edit and sign off every narrative
- » Owned capacity replaces per-token fees
- » Grow from 10 to 20 to 30 GPUs on the same host

Pinned, self-hosted models · one investigation agent

**4**

Alert triage

**4**

Evidence + narrative

**2**

Typology + QA

LIQID GPU chassis · 10 AMD Instinct™ MI350P · 1.44 TB HBM3E

PCIe Gen 5 connection · chassis to host

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

### Alert Triage · 4 GPUs

Llama 3.1 8B replicas score and route every alert.

**9.5x**

scaling, 1 to 10 GPUs

**44,383**

tok/s, 7B-class, 10 GPUs

### Evidence + Narrative · 4 GPUs

Llama 3.3 70B, one model per GPU, gathers evidence and drafts the SAR narrative.

**1,256**

tok/s, 72B-class, 8 GPUs

**1 GPU**

per 70B model at FP8

### Typology Check + QA · 2 GPUs

gpt-oss-120b tests each draft against known typologies.

**144 GB**

HBM3E per MI350P

**1 card**

holds gpt-oss-120b

Measured results: Liqid Phase-3 testing on 1 to 10 MI350P GPUs on one host. Test conditions and notes on page 2.

#### Built for

Financial crimes · AML / BSA · Fraud investigations · Model risk management

#### Proven on

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

**Size Your Alert Queue**

sales@liqid.com · +1 303.500.1551

## Proof: Measured on MI350P

No cluster: three models, ten GPUs, one server, and the next job lands on the same host.

| Result                                        | What was measured                                                                   | Why it matters here                                                                |
|-----------------------------------------------|-------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| <b>9.5× scaling</b><br>1 to 10 GPUs, 7B-class | 4,653 tok/s on 1 GPU to 44,383 tok/s on 10, one replica per GPU                     | Triage capacity grows with every GPU, with headroom to clear the backlog overnight |
| <b>1,256 tok/s</b><br>on 8 GPUs               | 72B-class model, FP8, under interactive limits: P99 TTFT ≤ 2 s, ≥ 30 tok/s per user | Narrative drafting at interactive speed for investigators on live cases            |
| <b>1 GPU</b><br>per 70B model                 | Llama 3.3 70B is ~70 GB at FP8 and fits on one 144 GB MI350P with room for cache    | Each step of the agent runs on its own GPUs, with no model split across cards      |
| <b>What it means</b>                          | <b>Every alert gets a first pass</b>                                                | <b>Investigators verify cited narratives instead of retyping them</b>              |

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. Llama 3.1 8B, Llama 3.3 70B and gpt-oss-120b performance for this layout is projected from these results.

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

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

|                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>10 GPUs</b><br>1.44 TB HBM3E<br><b>Financial crimes investigation agent</b><br>Triage, evidence, narrative and QA for every alert.<br><b>MODEL LAYOUT</b><br>4 × Llama 3.1 8B triage<br>4 × Llama 3.3 70B evidence and narrative<br>2 × gpt-oss-120b typology and QA<br><b>WHO BUYS, WHY NOW</b><br>Head of Financial Crimes / BSA Officer.<br>Alert backlog, exam findings, investigator attrition. | <b>20 GPUs</b><br>2.88 TB HBM3E<br><b>Bank-wide private model hub</b><br>One server for every line of business: triage, extraction, KYC, complaints, credit memos and research.<br><b>MODEL LAYOUT</b><br>1–2 GPUs per job from the 4-GPU briefs<br>6–8 GPUs of shared Llama 3.3 70B and gpt-oss-120b, interactive by day, batch by night<br><b>WHO BUYS, WHY NOW</b><br>CIO / Head of AI Platform. 5+ departmental pilots competing for GPUs; API bills growing. | <b>30 GPUs</b><br>4.32 TB HBM3E<br><b>Champion and challenger model validation</b><br>Production and challenger models side by side on identical inputs, plus frontier reasoning.<br><b>MODEL LAYOUT</b><br>8 × Mistral Large 3 (675B, FP8)<br>8 × challengers, e.g. Arcee Trinity Large (400B)<br>14 × workers (Llama 3.3 70B, Llama 3.1 8B)<br><b>WHO BUYS, WHY NOW</b><br>Chief Model Risk Officer / CRO. SR 11-7-style validation; board scrutiny of AI. |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Validation teams need several large models resident at once, which is a memory problem, not a compute one. Starting with one job? See the 4-GPU Classification and Triage 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.

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

Two questions settle the size of the opportunity: **how many alerts sit untouched today, and how long does each one take to close?**

### Bring Us Three Numbers

- 1 Alert volume:** alerts per month and the open backlog.
- 2 Case size:** documents and transactions per case.
- 3 Concurrency:** investigations working at peak.

**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. Outputs support, and do not replace, investigator decisions and SAR filing. AML spend: FIS, May 2026. Token data: Futurum Research, Sep 2026.