

The Autonomous AI Trading Node

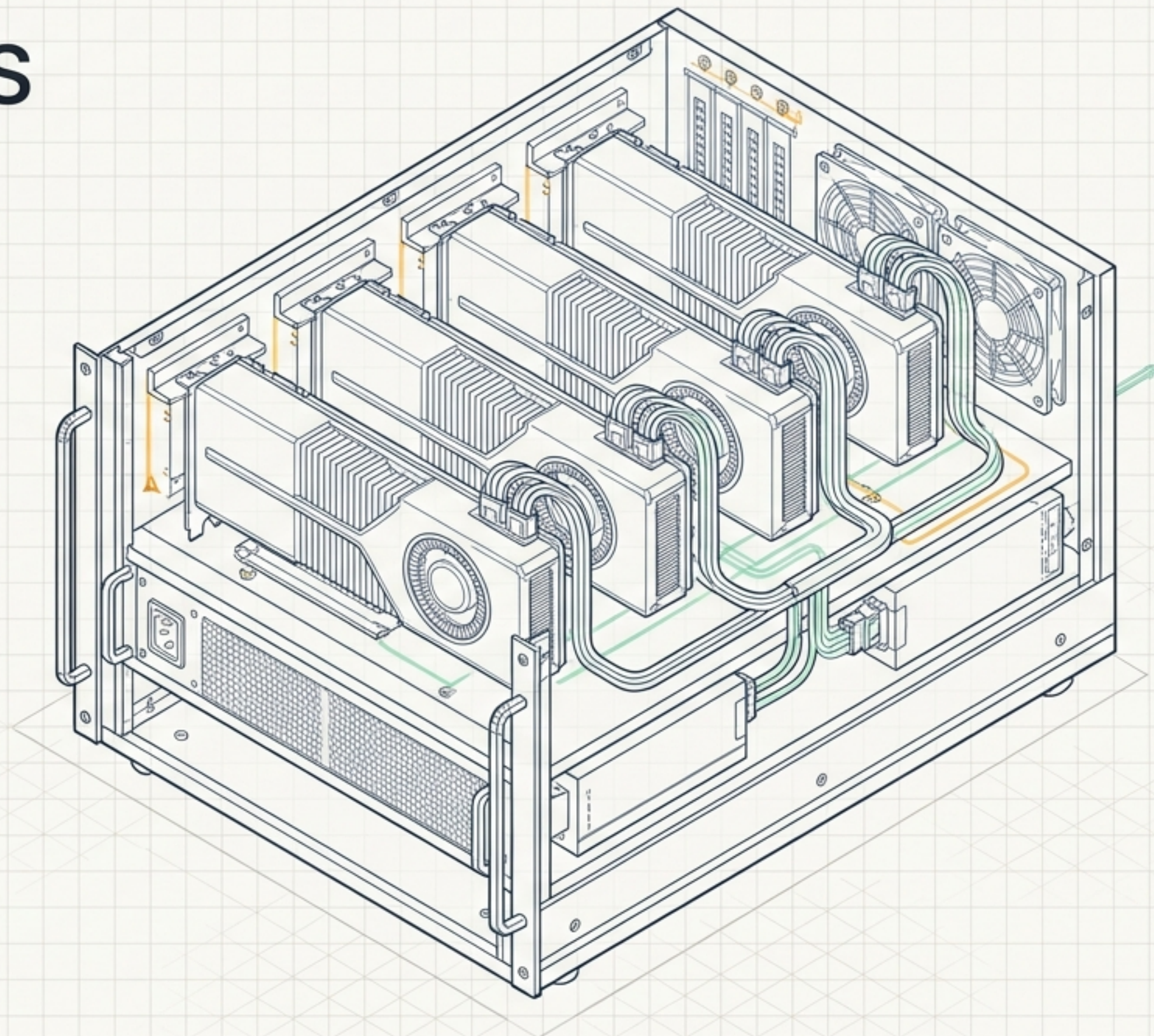
Building a Localized, Institutional-Grade Algorithmic Alpha Engine.

System Specifications:

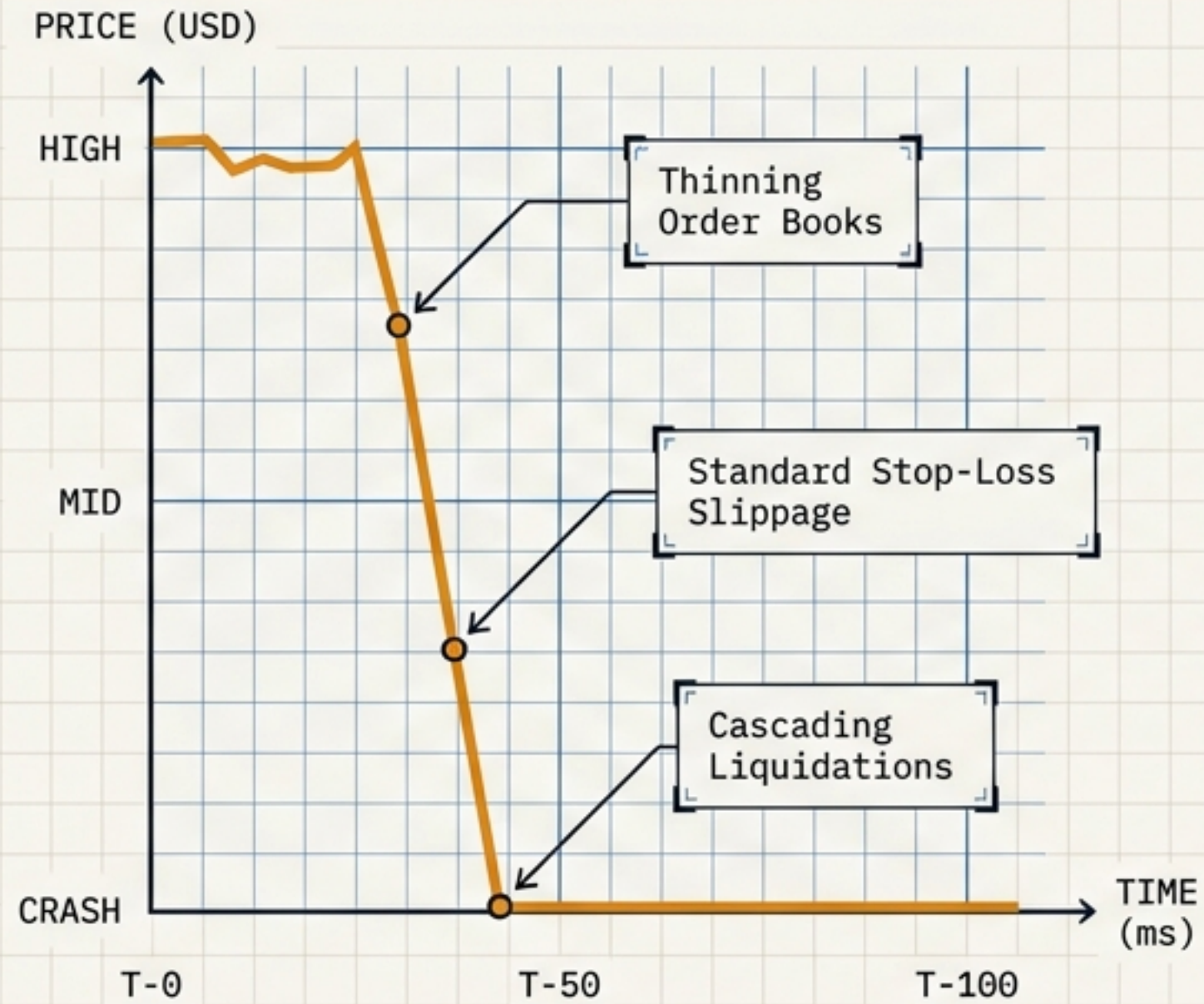
Total VRAM: 64GB GDDR7

Network: Tri-WAN Redundancy

Architecture: Dual-Brain (Rust/Node)

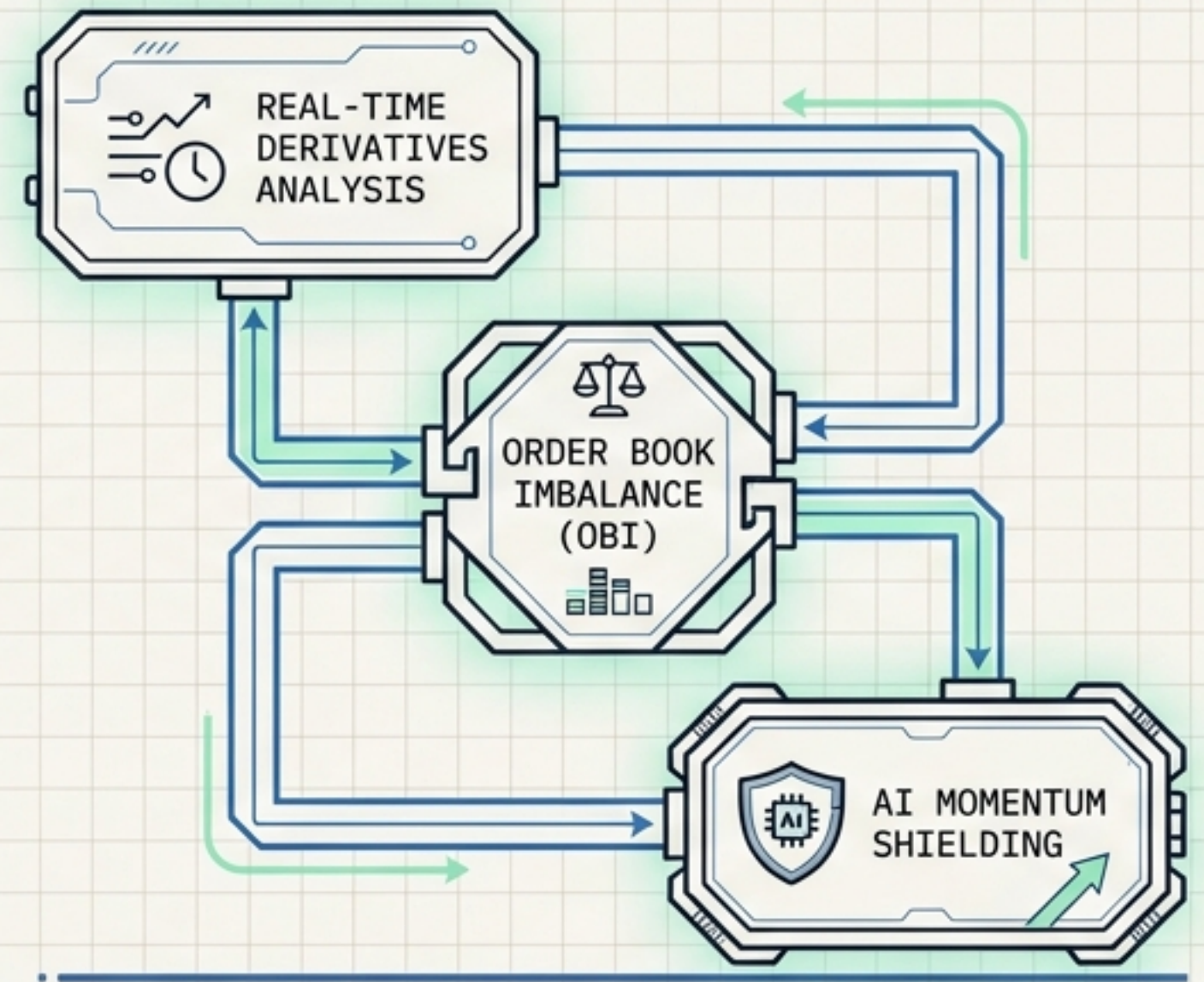


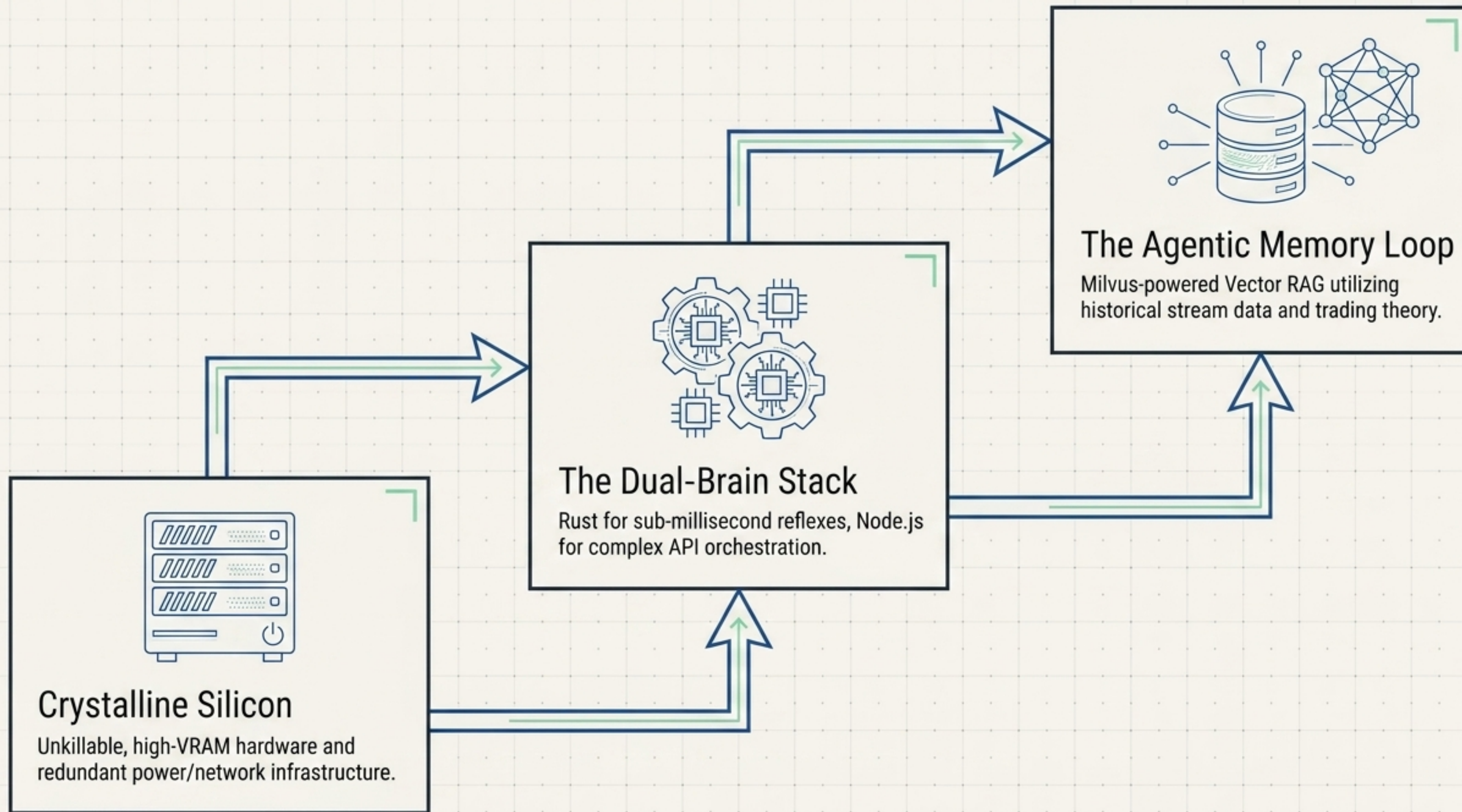
THE 2022 VULNERABILITY



Lagging spot prices are fatal. Survival requires reacting to derivative metrics and order flow before the spot price crashes.

THE MOMENTUM SHIELD





Finding the VRAM Sweet Spot

	2x Intel Arc Pro B70	AMD Ryzen AI Max 395	4x RTX 5060 Ti
Total VRAM	64GB VRAM	Up to 128GB Unified	64GB VRAM
Memory Bandwidth	High	256 GB/s (Slower)	448 GB/s (GDDR7)
Software Ecosystem	SYCL/oneAPI (Experimental bugs)	ROCm	Flawless CUDA stability
Estimated Price	~26,000 SEK	~46,000 SEK	~41,200 SEK

The undisputed winner for local inference speed and CUDA stability.

The 64GB VRAM Open-Frame Build

CPU: AMD Ryzen 9 7900X

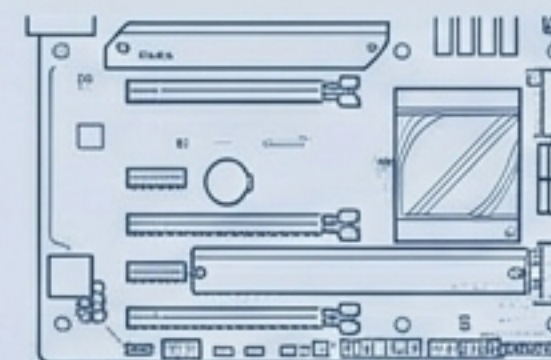
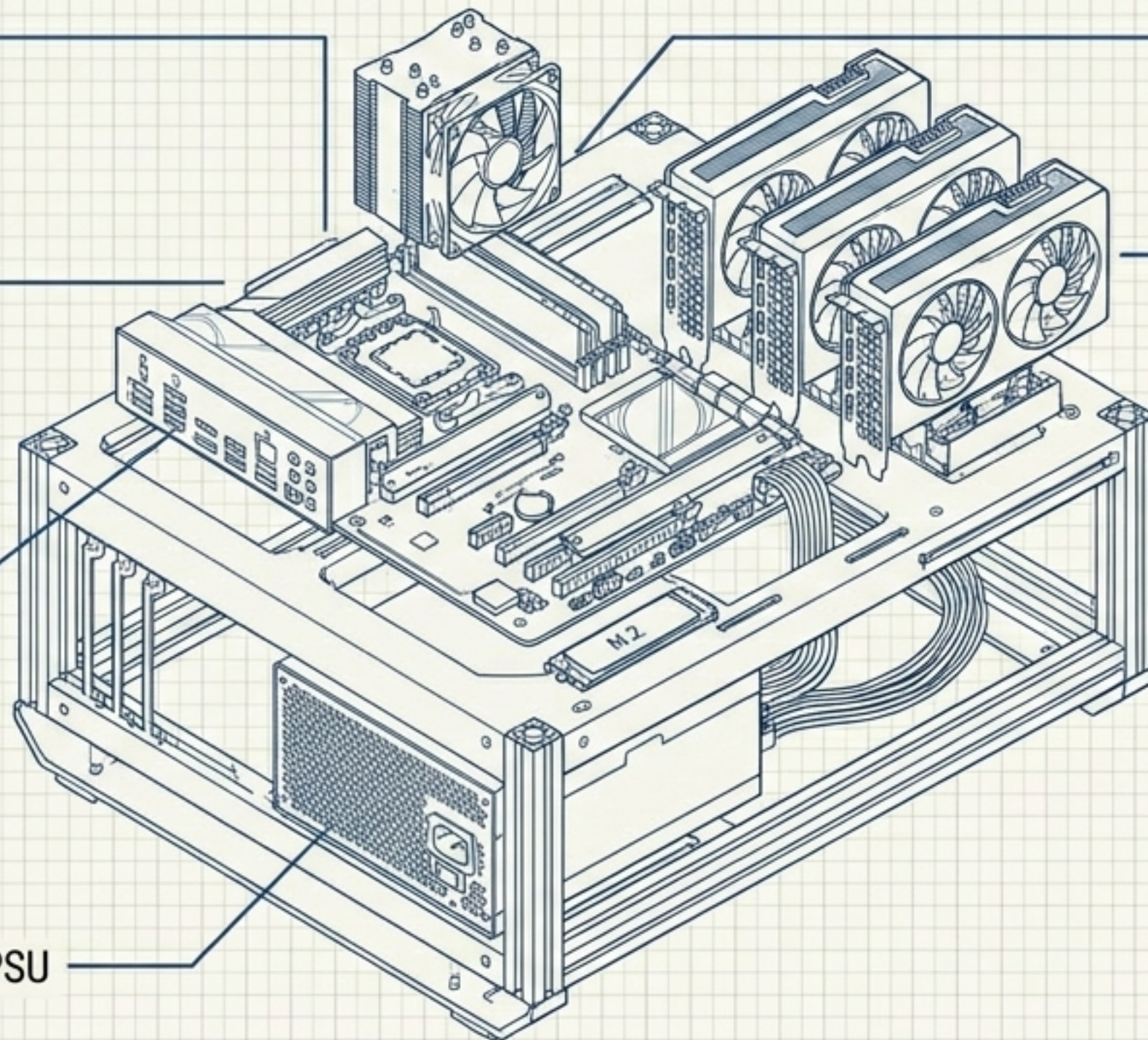
Motherboard: ASUS ProArt
X670E-Creator WiFi

RAM: 96 GB DDR5
6000MHz CL30

Power: 1600W Tier-A ATX 3.0 PSU

RAM: 96 GB DDR5 6000MHz CL30

GPUs: 4x Gigabyte RTX 5060 Ti 16GB



The Lane Allocation Trick

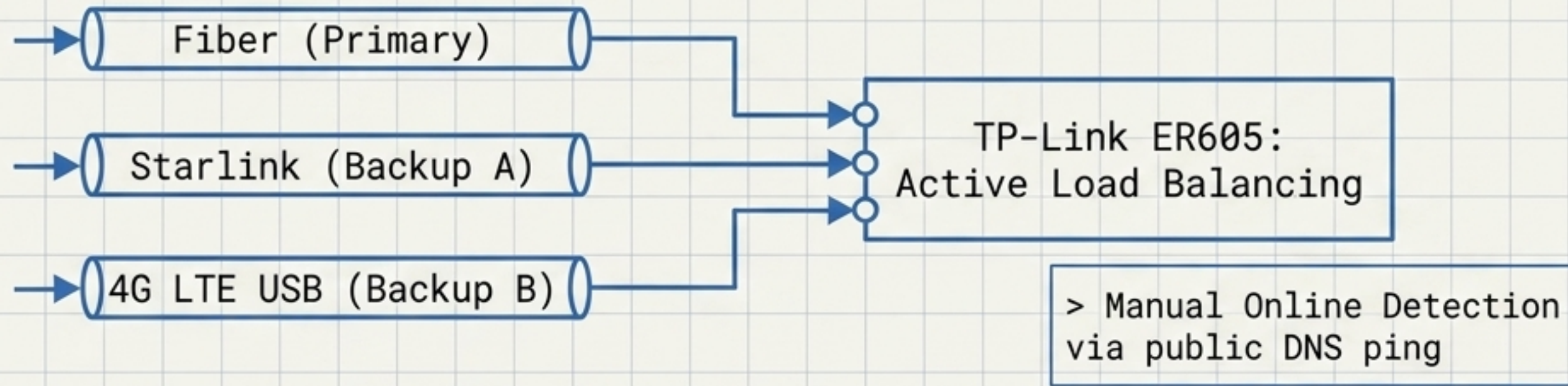
Top slots: PCIe 5.0 x8 / x8

Bottom slot: PCIe 4.0 x4

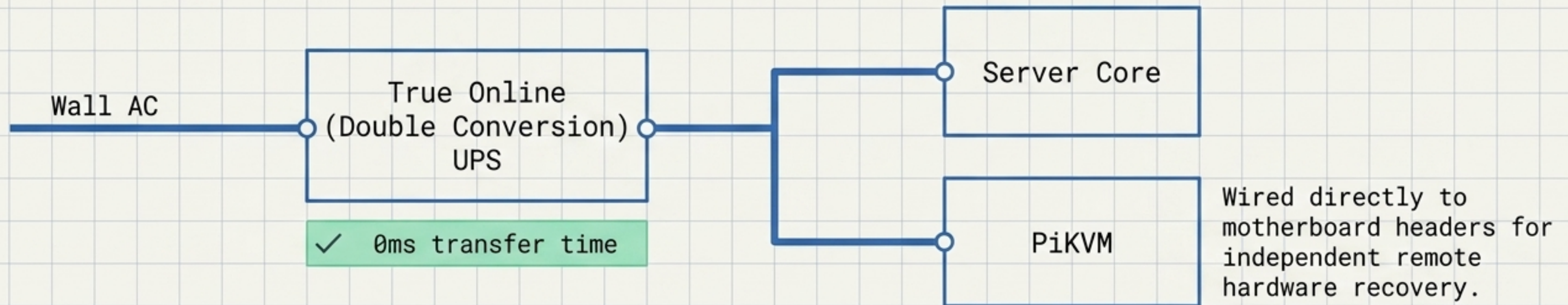
4th GPU: M.2-to-PCIe x4 lane adapter

Triple-Redundancy Network & Power

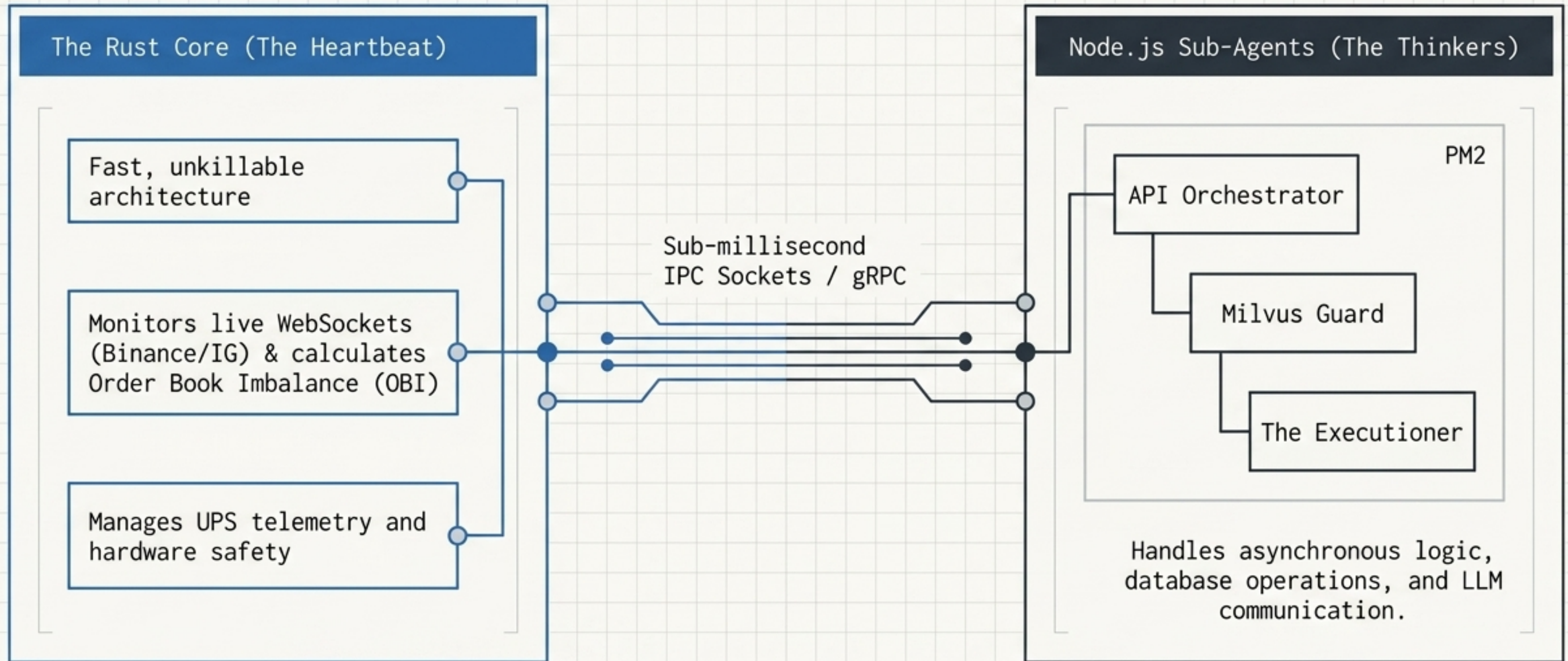
Network Redundancy Flow



Power & Control Schematic



Architecture of the Dual Brain



The Local AI Engine

Model Role Matrix

The Executive (Tool Calling)

Model: `Qwen 3.6 Coder (35B)`

Virtually flawless JSON output and precise API execution.

The Analyst (Long-Context)

Model: `Gemma 4 (31B)`

Massive 128k+ context window. Ideal for digesting hours of historical tick data.

The Reasoner (Risk Logic)

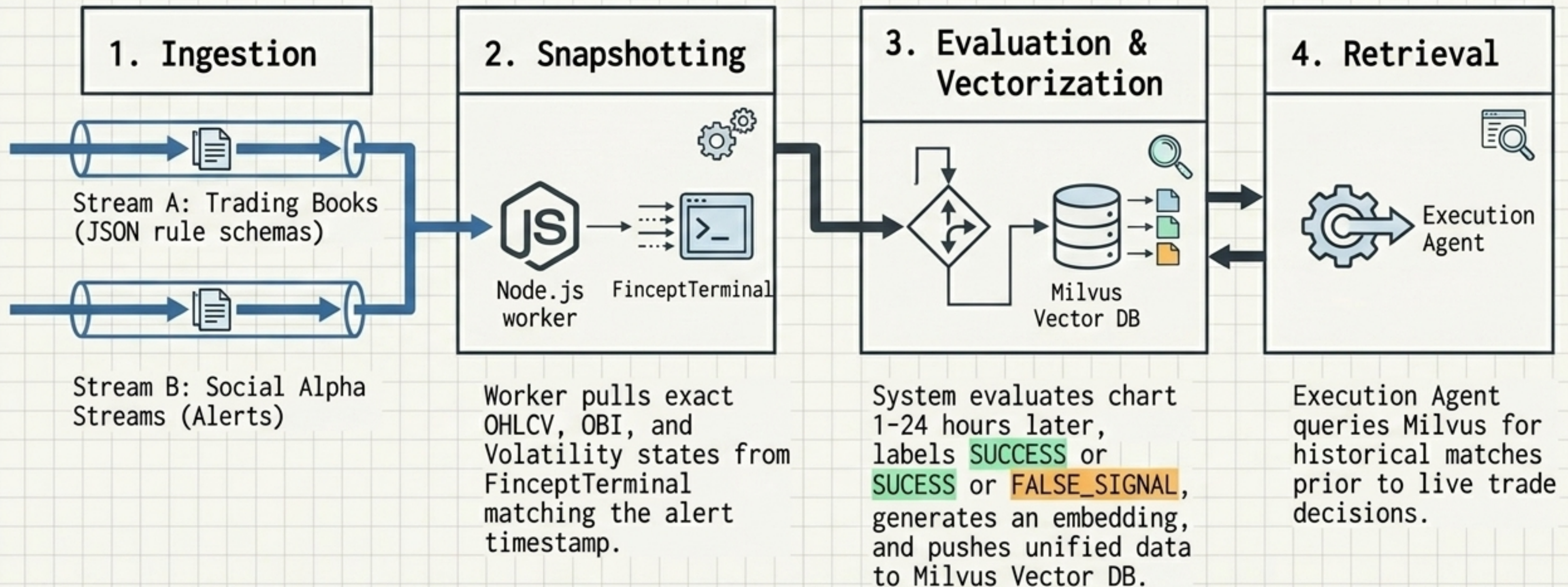
Model: `Mistral Small 4`

Deep dependency-checking and logic traces before trade execution.

```
$ llama-server --model qwen.gguf -ngl 99 --flash-attn compiled with GGML_CUDA=ON
```

Seeding the Knowledge Base (Milvus)

Input/Output



The Knife-Catcher Protocol (Execution Logic)



Point A: The Fall

OBI drops below threshold.
Rust bypasses the LLM and instantly
sends an IPC command to Node to execute
an aggressive Post-Only Short.

Point B: The Rebound

Open Interest stops dropping, buy blocks appear.
Node queries local state for margin safety,
then asks Qwen for a rapid Milvus sanity-check
before placing the buy order.

The Chaos Monkey Stress Test

Capital scales only when the system survives these synthetic failures.



1. Network Severance

Ensure ER605 active load balancing routes WebSocket traffic to Starlink/4G without dropping active connections.



2. Process Assassination

Kill a Node execution process mid-stride. Verify the Rust core intercepts the crash, shields capital, and commands PM2 to reload the worker.

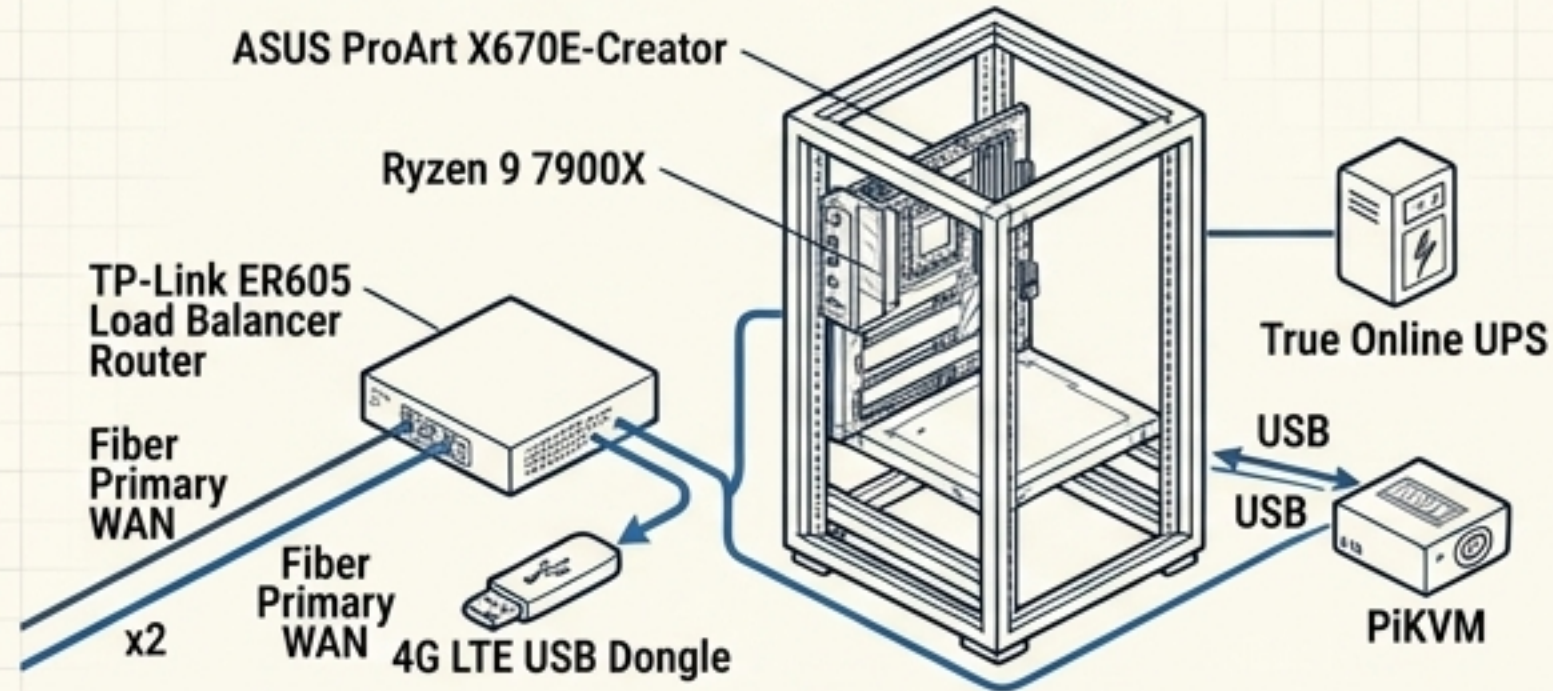


3. Grid Blackout

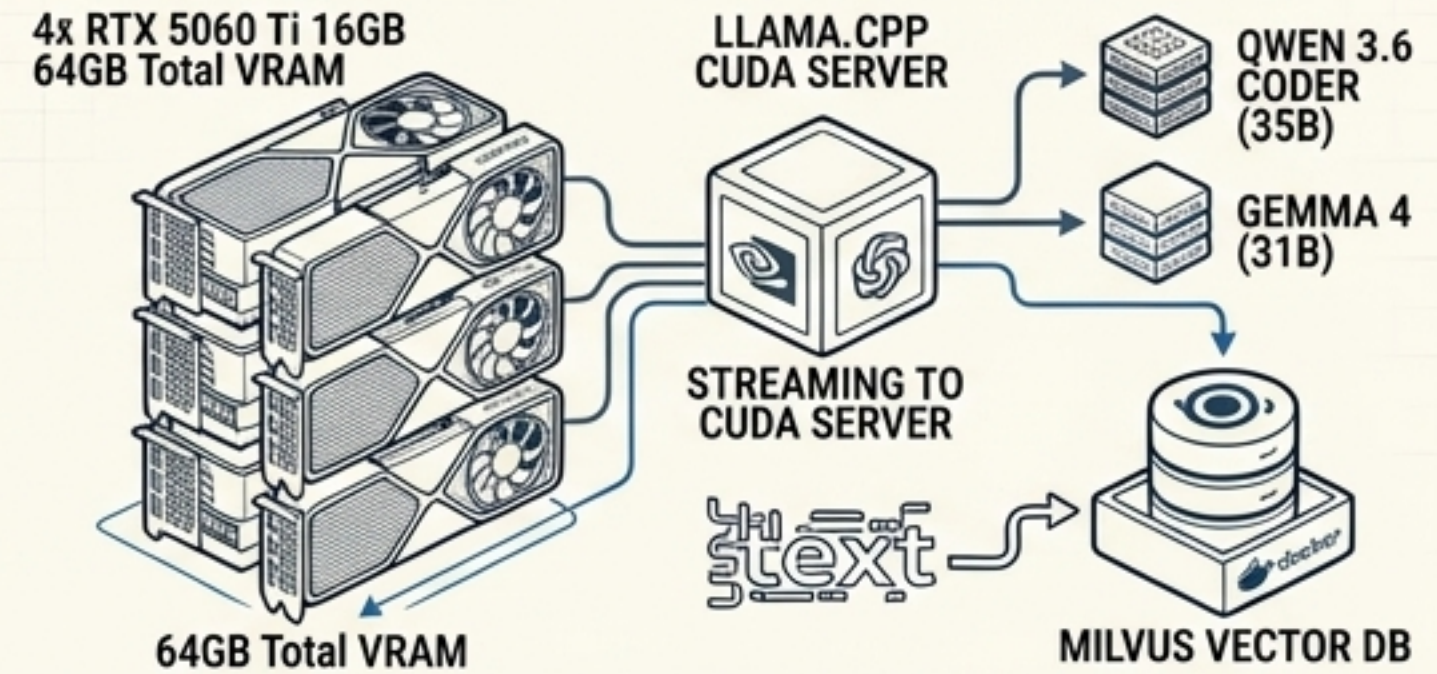
Cut server room power. Verify True Online UPS 0ms transfer and ensure Rust fires emergency API calls to flatten open positions before safe shutdown.

Deployment Roadmap

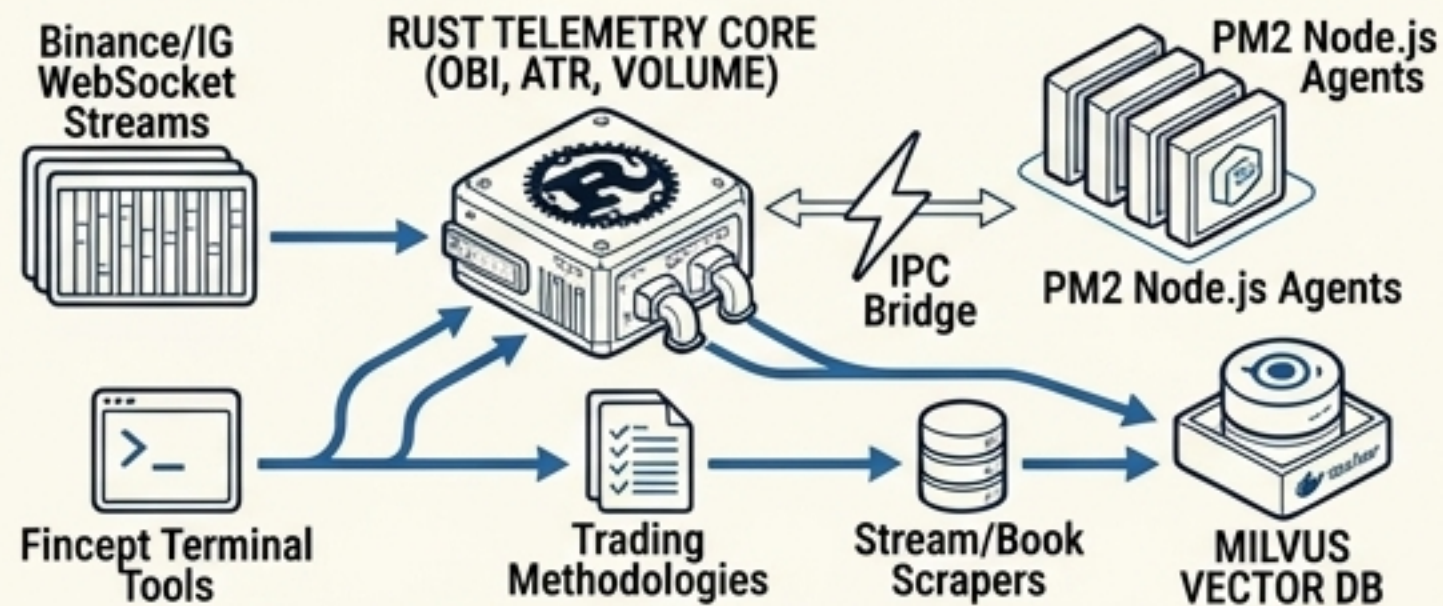
PHASE 1: INFRASTRUCTURE & NETWORK DEPLOY



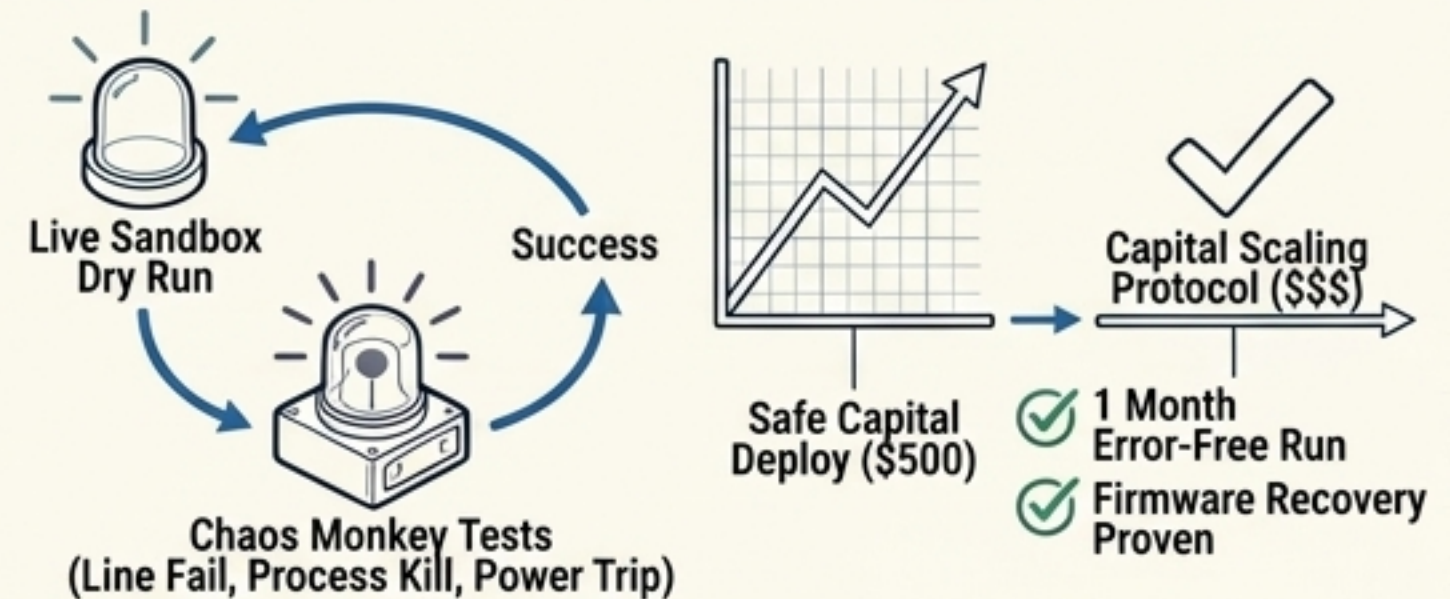
PHASE 2: AI CORE, EMBEDDING, & MILVUS



PHASE 3: TELEMETRY CORE & ALPHA HARVESTING



PHASE 4: RISK SANDBOX & CAPITAL SCALING



Phase 1 (Days 1-5):
Infrastructure & Network Deploy
(Bare metal build, OS, ER605 active load balancing).

Phase 2 (Days 6-10):
AI Core, Embedding, & Milvus
(llama.cpp compilation, downloading Qwen).

Phase 3 (Days 11-17):
Telemetry Core & Alpha Harvesting
(Rust heartbeat loop, Fincept feeds, Milvus seeding).

Phase 4 (Days 18-30):
Risk Sandbox & Capital Scaling
(Paper trading, Chaos monkey tests, real capital deployment).