

RALFIIA HYBRID OPS COPILLOT

Sovereign Edge-Cloud Orchestration & ROCm Hardware-Aware Auditing

- 🛡️ Sovereign gateway routing enterprise tasks.
- 🔧 Edge optimization on AMD Ryzen (\$0 cost).
- ☁️ On-demand cloud scale via AMD Instinct MI300X.



THE HIDDEN COST

Cloud Dependency & Inefficiency

Bleeding Margins

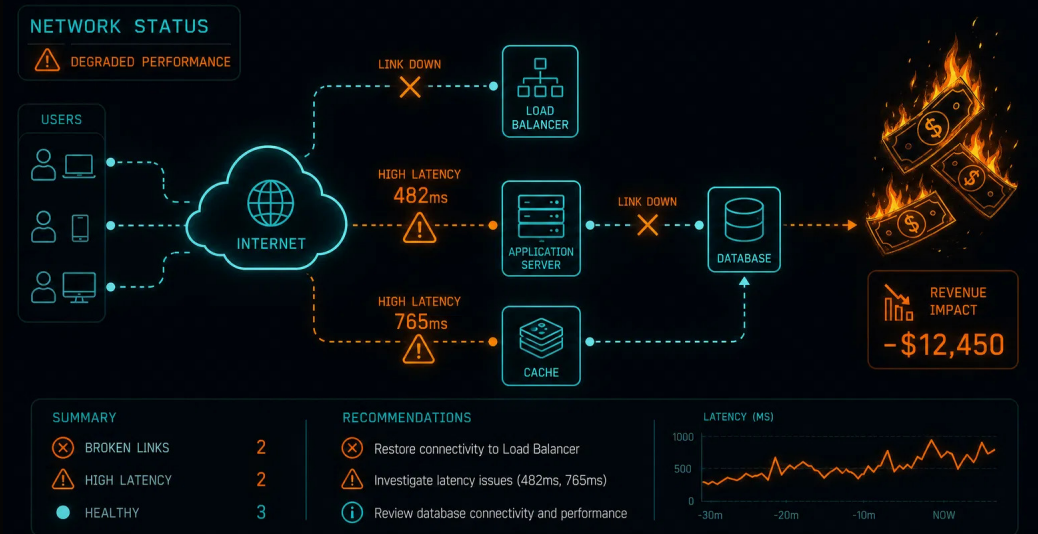
Over-reliance on public LLM APIs wastes budget on trivial tasks that could be handled locally.

SRE Blindspots

Lack of secure, localized self-healing scripts for real-time infrastructure repair and monitoring.

Latency Trap

High delays when generating complex technical quotes via monolithic, distant cloud models.



A2A MESH

Agent-to-Agent Mesh Infrastructure

Root Gateway

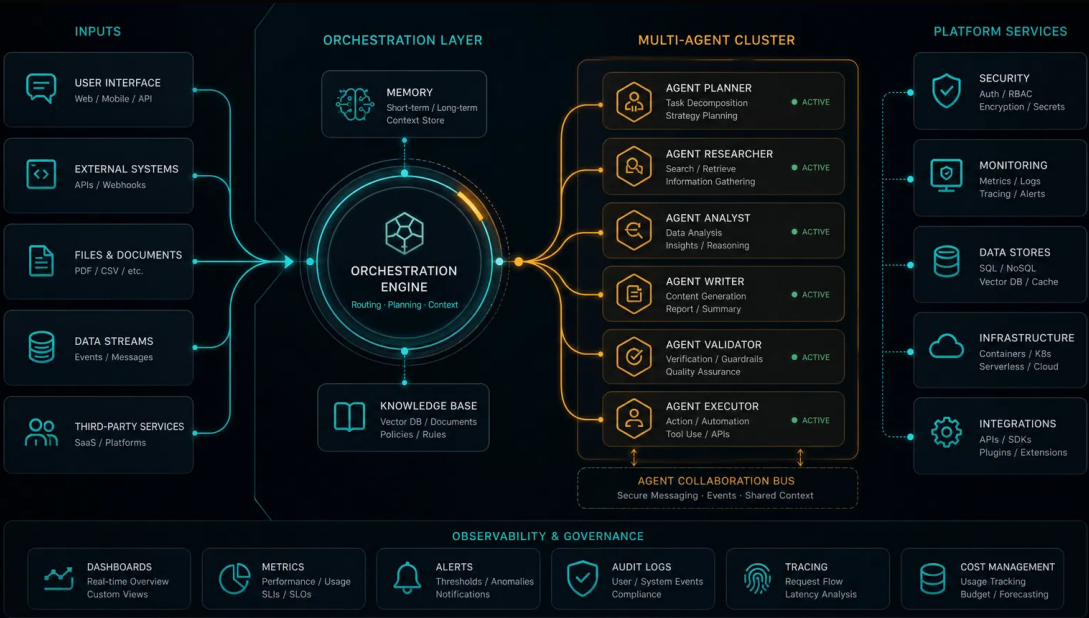
Sovereign orchestrator evaluating syntactic complexity in milliseconds to route tasks efficiently.

CrewAI Smart Quoter

Specialized agents generating and validating complex tech budgets locally on edge hardware.

SRE Watchdog

LangGraph cyclic state agent automatically fixing infrastructure code failures in real-time.



SOVEREIGN HYBRID ROUTING

Edge Efficiency meets Cloud Scale

EDGE

AMD Ryzen Optimization

Processes low-complexity tasks locally.

Operational Cost: **\$0.00**

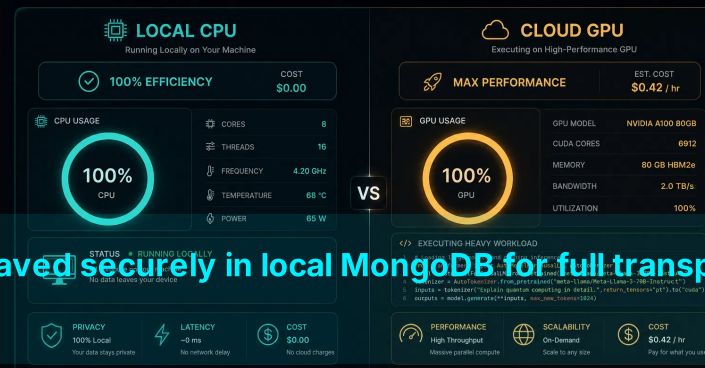
CLOUD

AMD Instinct MI300X

Processes complex logic & code generation via Fireworks.

On-Demand Scaling

 **IMMUTABLE AUDITING:** Token telemetry saved securely in local MongoDB for full transparency.



Choose the right compute for your workload: Local for privacy & zero cost • Cloud for maximum performance & scale

ROCM AUDITOR

Hardware-Aware Security & Validation

ROCm Interception

Code-level interceptor inspecting developer scripts in real-time to ensure **sovereign compliance**.

VRAM & HIP Validation

Evaluates memory limits and automatically ports **CUDA kernels** to AMD HIP for seamless execution.

Zero-Waste Deployment

Prevents deployment runtime errors before code hits physical **AMD Instinct** hardware.

AMD INSTINCT™ MI300X

LEADERSHIP PERFORMANCE. OPEN ECOSYSTEM.

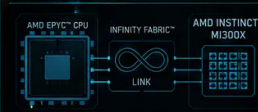
ARCHITECTURE

CDNA™ 3
192 GB HBM3
5.3 TB/s MEMORY BANDWIDTH
182.0 TFLOPS FP16
91.0 TFLOPS FP64
FP8 / FP16 / BF16 / FP64 SUPPORT
AMD INFINITY FABRIC™

SPECIFICATIONS

PROCESS	5nm
DIE SIZE	814 mm²
TRANSISTORS	153.6B
HBM3 MEMORY	192 GB
MEMORY BANDWIDTH	5.3 TB/s
TDP	750W

SYSTEM TOPOLOGY



PERFORMANCE

FP16	182.0 TFLOPS	██████████
FP8	364.0 TFLOPS	██████████
BF16	182.0 TFLOPS	██████████
FP64	91.0 TFLOPS	██████████

HIP
COMPLIANT

- ✓ HIP API SUPPORTED
- ✓ KERNEL COMPILEATION
- ✓ MEMORY MANAGEMENT
- ✓ DEVICE COMPATIBILITY
- ✓ PERFORMANCE OPTIMIZED
- ✓ PORTABLE CODE READY

STATUS: COMPLIANT

PRODUCTION READY

Enterprise-Grade Deployment Architecture



MULTI-PLATFORM

Compressed linux/amd64 Docker image deployed on ghcr.io for seamless cross-environment compatibility.



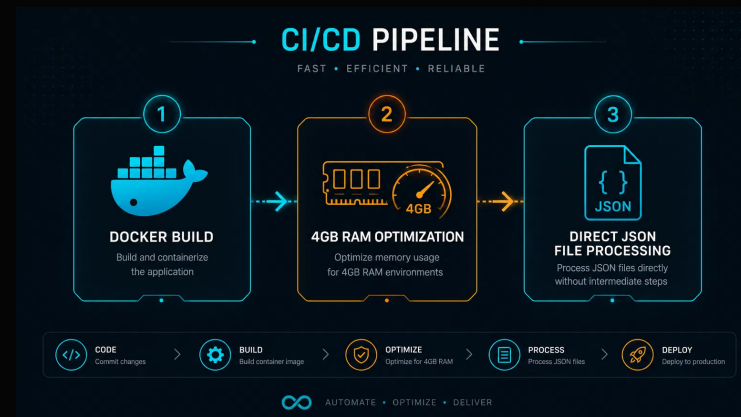
4GB RAM LIMIT

Ultra-lightweight runtime optimized for automated evaluation and high-density edge environments.



SYNC I/O PIPELINE

Millisecond processing reading /input/tasks.json and exporting direct results to /output/results.json.



SOVEREIGN VISION

Democratizing AI for Latin America

Engineering Leadership

Developed by Team InnerSpark / PC Doctor (Rafael Lopez). Expert-led AI architecture.

Local Focus

Enterprise AI architecture optimized for real-world clients and infrastructure in Latin America.

Digital Sovereignty

Empowering local businesses with self-hosted, high-performance AI orchestration.

