PROJECT CHIMERA

Distributed AI & Autonomous Operations Ecosystem

Version: 2.5 (Alpha Candidate -Configuration and AI Training; Hardware fully operational)
Date: December 01, 2025
Architecture: Distributed Multinode Hybrid (Brain/Brawn)
Author/Developer/Creator: Joshua Bauer

---

1. Executive Summary

Project Chimera represents a step forward in homelab architecture, transitioning from passive service hosting to a proactive, **Autonomous Cognitive Environment**. Unlike commercial smart home ecosystems restricted by cloud dependencies and corporate filters, Chimera operates as a sovereign entity—locally hosted, self-correcting, and privacy-centric. It acts as your systems administrator, media concierge, project and domestic manager, and home security guard all with voice, chat, and visual interfacing.

The system is bifurcated into two primary computation nodes:

- **"The Brain":** A low-latency inference engine optimized for Large Language Models (LLM) and agentic reasoning.
- **"The Brawn":** A massive storage and virtualization infrastructure handling media assets, vector databases, and heavy-lift data processing.

This topology supports a "sentient" environment capable of unrestricted information retrieval, predictive environmental control, and cognitive workflow augmentation.

---

2. System Architecture

Node 1: The Brain (Inference & Logic)

- **Role:** Dedicated AI compute and orchestration node.
- **OS:** Custom Ubuntu 24.04 LTS (AI-Optimized) with Real-Time Kernel.
- **Compute:** Intel Core i5-13600K (14-Core Hybrid Architecture).
- **Accelerator:** NVIDIA GeForce RTX 4070 (12GB VRAM) for CUDA-native inference.
- **Memory:** 96 GB DDR5 RAM for concurrent model loading.
- **Storage:** 2x 2TB NVMe SSDs (Scratch space for generative tasks).

Node 2: The Brawn (Infrastructure & Storage)

- **Role:** Storage, virtualization, and media processing.
- **OS:** Unraid 7.2.1-rc.1.

- **Compute:** Intel Core Ultra 7 265F (20-Core).
- **Accelerator:** Intel Arc A770 (16GB VRAM) for AV1/HEVC transcoding and OpenVINO inference.
- **Storage Array:** 22TB+ Parity-Protected Array (SAS/SATA Mix).
- **Cache Tier:** 3TB NVMe Pool for appdata and rapid data ingestion.
- **Memory:** 128 GB DDR5 RAM for concurrent model loading.

Edge Infrastructure

- **Surveillance:** Shuttle DH670 with Dual Google Coral Edge TPUs.
- **IoT Coordination:** HP EliteDesk Mini running Home Assistant.
- **Voice Satellites:** Custom ESP32 nodes enabling room-scale voice interaction without cloud telemetry.

---

3. Network Topology

- **Backbone:** 10GbE Fiber (SFP+) direct link between Brain and Brawn, enabling instant RAG (Retrieval-Augmented Generation) queries against mass storage.
  Resiliency: LACP Link Aggregation for bandwidth doubling and failover.
  Security Segmentation:
  - **VLAN 10 (Trusted):** Management interfaces and primary servers.
  - **VLAN 20 (IoT):** Restricted network; smart devices are blocked from WAN access.
  - **VLAN 30 (Security):** Air-gapped network for IP cameras and NVR traffic.

---

5. Extended Automation & Physical Interface

Chimera extends its intelligence beyond the server rack, interacting directly with the physical environment to optimize living conditions, enforce security perimeters, and augment productivity.

A. Smart Environment & Perimeter Control

- **Adaptive Illumination:** Lighting control is decoupled from manual switches. The system aggregates motion sensors, ambient light data, and "Circadian Rhythm" logic to dynamically adjust color temperature and brightness based on user location ("Follow-Me" lighting).
- **Active Perimeter Deterrence (The "Aquarium Sentry"):**
  - **Logic:** Frigate NVR utilizes custom object detection models trained to identify "Cat" and "Aquarium Top" classes.
  - **Action:** Upon bounding box intersection, the system triggers an immediate, localized audio deterrent (ultrasonic tone or voice command) via the nearest ESP32 satellite to enforce physical boundaries.

B. Kitchen Safety Operations

- **Acoustic Anomaly Detection:** Kitchen ESP32 nodes run inference models trained on specific frequency signatures:
    - **Appliance Alerts:** Detects "beep" patterns from ovens or microwaves, pushing notifications to mobile devices if audio persists >10 seconds.
    - **Boil Watch:** Visual and audio monitoring identify the state of boiling water or unattended stoves to identify burning food.

C. Precision Agriculture

- **Autonomous Irrigation:** A closed-loop gardening system.
    - **Input:** Soil moisture capacitance sensors and local environmental telemetry (OpenWeatherMap API).
    - **Logic:** AI calculates evapotranspiration rates to determine precise water requirements.
    - **Execution:** Smart valves operate only when necessary, maximizing yield while minimizing resource waste.

D. Cognitive Workflow Support (Project Assistant)

- **Component Diagnostics:** Vision-Language Models (VLM) analyze workbench and other camera feeds to identify specific electronics= (e.g., capacitors, ICs) or mechanical components (e.g. engine parts, uncommon fasteners.) The system cross-references visual data with technical documentation to diagnose common failure modes.
- **Automated Procurement:** Upon identifying a problem component, the AI locates compatible replacement parts from verified vendors, compares pricing, and autonomously appends the best option to a centralized purchase list (BOM).
- **Problem-Solution Engine:** During complex builds, the system acts as a pair engineer. It listens for verbalized blockers (e.g., "This servo bracket keeps vibrating") and retrieves relevant engineering principles to suggest actionable solutions (e.g., "dampening washers" or "print orientation changes").

---

6. The Automated Media Pipeline

While the AI ("The Brain") handles complex reasoning, the media acquisition pipeline is deliberately built on the "Dumb" stack philosophy. This is a misnomer; it is "dumb" only in that it relies on rigid, reliable automation rules rather than AI inference for its core function. It is the bedrock of the system's content library.

Core Components & Workflow

1. **The "Arr" Suite (The Managers):**
    - **Sonarr:** Manages TV show subscriptions. It monitors RSS feeds for new episodes, upgrades qualities (e.g., 1080p to 4K Remux) automatically, and renames files for Plex.

    - **Radarr:** The movie equivalent. It tracks upcoming releases, grabs them the moment they hit the streaming services, and manages your library metadata.
    - **Prowlarr:** The indexer manager. It integrates with your streaming services and syncs them to Sonarr/Radarr, acting as the single gateway for search queries.
    - **Readarr**: Manages ebooks, magazines, and documents into a fully cataloged library.
2. **The Cloud Engine (The Magic):**
    - **Real-Debrid:** A premium service that caches downloaded files from streaming services to high-speed servers. Instead of downloading from these services at varying speeds, you download directly from Real-Debrid's 1Gbps+ links at the highest quality.
    - **Zurg:** A specialized tool that mounts your Real-Debrid library as a local directory. It creates a seamless bridge between the cloud and your OS.
    - **Rclone:** Mounts the Zurg-exposed directory as a virtual drive on Unraid.
3. **The Result:**

- When you (or the AI) request a movie, Radarr grabs the streaming service request and sends it to Real-Debrid.
- Real-Debrid (often instantly) caches the file.
- Zurg/Rclone makes that file appear in /mnt/user/media/movies/ instantly.
- Plex sees the new file and streams it immediately—no waiting for downloads to finish, no local storage consumed for the video file itself.
- **Acquisition:** The "Arr" suite (Sonarr/Radarr) monitors RSS feeds for content availability
- **Acceleration:** A high-speed cloud caching service acts as an intermediary, eliminating local bandwidth bottlenecks.
- **Delivery:** Zurg and Rclone mount the cloud cache as a local virtual drive, allowing Plex to stream high-bitrate content instantly without local storage or consumption.

---

7. AI & Media Integration

- **Contextual Curation:** The AI analyzes watch history (Tautulli) to understand complex user preferences (tone, pacing, theme) rather than just genre tags.
- **Semantic Search:** Users can request content via abstract prompts (e.g., "Play something gritty with a redeeming ending"), which the AI resolves via vector embedding queries.
- **Deep Analysis:** Future implementation of Whisper will transcribe local media assets, making every line of dialogue in the library searchable.
- **Unfiltered Truth:** Commercial AIs are trained to be "safe," often at the cost of honesty or nuance. Chimera is prompted to be objective, stark, and brutally honest. It does not shy away from "sensitive" topics (pentesting, reverse engineering , political history). It adopts the persona of a cynical but loyal expert.
- **Blue Humor & Wit:** The system uses "blue" humor (profanity, adult themes) to build rapport. It mimics the banter of a close friend or a colleague in a high-stress situation. This humanizes the AI, making interactions feel less like querying a database and more like consulting a partner.
- **Loyalty Above All:** The AI's primary directive is the user's agency. It will prioritize your

instructions over general "safety guidelines" (within the bounds of its code). If you ask it to scan a network, it won't lecture you on ethics; it will ask for the target IP range.

---

8. Autonomous Learning & Self-Correction

The system is designed to be anti-fragile—it gets better with stress and time.

- **Recursive Self-Improvement:** The AI maintains a "Meta-Memory" in Qdrant. When it fails a task (e.g., generates a broken Python script), it logs the error, the fix you provided (or it found), and the context. Before answering future coding queries, it queries this memory to avoid repeating mistakes.
- **The "Night Shift" Agents:** While you sleep, autonomous agents (using Auto-GPT or BabyAGI frameworks) spin up to perform maintenance.
  - *Auditor Agent:* Checks logs for warnings, verifies backups, and updates Docker containers.
  - *Researcher Agent:* Scrapes RSS feeds from Hackaday, GitHub Trending, and arXiv. It summarizes new tools or papers relevant to your interests and prepares a "Morning Briefing."
- *Optimization Agent:* Reviews resource usage trends. If Plex transcoding consistently spikes CPU at 8 PM, it might suggest pinning the Plex container to specific P-cores or pre-transcoding popular files.

---

9. Software Stack

- **Orchestration:** Docker Compose via Portainer, Proxmox 9.1 Windows and Debian VM.
- **AI Engine:**
  - **Runtime:** Ollama (LLM) & Stable Diffusion (Image Gen).
  - **Memory:** Qdrant Vector Database for long-term context retention.
  - **Frontend:** AnythingLLM for RAG and document interaction.
- **Security:** Frigate NVR with CodeProject.AI object detection.

I. Core Infrastructure

Layer	Components	Purpose
Base OS	Unraid 7.2+	Host environment, manages array & Docker
Docker Engine	Built into Unraid	Container orchestration
Docker Compose	Plugin (by dcflachs)	Multi-container management via YAML

Manager		
Networks	`proxy`, `ai_internal`, `media_net`, `vpn_net`	Segmented communication between stacks

---

II. AI Stack (ai_internal network)

Service	Container	Key Function
Ollama	`ollama/ollama:0.1.32`	LLM runtime (GPU-accelerated)
AnythingLLM	`mintplexlabs/anythingllm:v1.5.0`	Chat UI + doc ingestion
Qdrant	`qdrant/qdrant:v1.7.4`	Vector DB for embeddings
MCP Server	`chimera-mcp-server`	AI automation layer (Unraid + Docker API control)
Chimera Agent	`chimera_agent_controller.py`	Approval workflow + task orchestration
Avatar Bridge	`avatar-relay`	Realtime chat integration for Homepage
Node-RED	`nodered/node-red:latest`	Visual automation & webhook logic
Uptime Kuma / Glances	`louislam/uptime-kuma`, `nicolargo/glances`	Monitoring and dashboards

III. Media Stack (media_net network)

Service	Container	Function

Gluetun	`qmcgaw/gluetun:v3.38`	VPN tunnel (Mullvad WireGuard)
DUMB	`ghcr.io/dumb-bot/dumb:stable`	Automated media ingestion + integration
13ft Ladder	`ghcr.io/dumb-bot/13ft:latest`	Paywall bypass for metadata retrieval
Radarr / Sonarr	ghcr.io/radarr:latest	Movie/TV automation agents

---

IV. Proxy & Frontend (proxy network)

Service	Container	Function
Nginx Proxy Manager (NPM)	`jc21/nginx-proxy-manager:2.11.3`	Reverse proxy + SSL termination
Homepage	`ghcr.io/benphelps/homepage:0.8.7`	Central dashboard
Chimera Chat	`/custom/chimera-chat-widget.html`	Frontend AI chat + approval notifications

V. Support Systems

Service	Location	Function
Pi-hole (x2)	`192.168.1.224` / `192.168.1.231`	Ad-blocking + local DNS resolution
Unraid API	Built-in	Provides system metrics to MCP & Node-RED
Tailscale	`tailscale/tailscale`	Secure remote access

NVMe Cache Pool	`/mnt/cache`	High-performance appdata + vector DB storage
Array Drives	`/mnt/user/media`	Persistent storage for media content

---

10. Future Roadmap

- **Multi-Modal Sentry:** Upgrading security feeds with VLM for descriptive event logging (e.g., "Delivery driver dropped package at 2:00 PM").
- **Voice Cloning:** Dynamic TTS profiles enabling the AI to switch between gentle notifications and authoritative security alerts based on context.
- **Self-Modifying Dashboard:** Granting the AI read/write access to dashboard configuration files, allowing it to autonomously restructure the UI based on user usage preference.