CommonCent: Sovereign LLM Hosting Architecture

This guide outlines how to leverage your dual-machine setup to create a private, high-performance financial analysis environment.

1. Hardware Allocation Strategy

Component	Machine: Desktop (Inference Server)	Machine: M3 Max (Client & Training)
Role	24/7 API Server	Development, UI, & Large Context Processing
Specs	RTX 3080 (10GB), Ryzen 7, 32GB RAM	M3 Max (36GB Unified Memory)
Primary Task	Rapid News Sentiment & SEC Flux Analysis	Fine-tuning & Portfolio Optimization
Constraint	VRAM Limited (10GB)	Compute speed lower than NVIDIA

2. Server Setup (The Desktop)

To make your desktop resources accessible to your remote laptop privately, we will use **Ollama** paired with **Tailscale**.

Step A: The Inference Engine (Ollama)

1. **Install Ollama** on your Windows/Linux desktop.
2. **Environment Variable:** Set OLLAMA_HOST=0.0.0.0 to allow connections from your private network.
3. **Model Selection:** * **Llama-3-8B (Quantized):** This will fly on your 3080, taking up ~5.5GB of VRAM, leaving room for system overhead.
    - **Mistral-Nemo-12B:** At 4-bit quantization, this provides higher reasoning for complex financial "Flux" explanations while fitting in your 10GB VRAM.

Step B: The Private Tunnel (Tailscale)

Instead of port forwarding (which is risky), install **Tailscale** on both machines.

- It creates a "MagicDNS" name for your desktop (e.g., desktop-server).
- You can now reach your LLM from your M3 Max anywhere in the world using: http://desktop-server:11434.

3. The Localized Intelligence Pipeline

The "Analyst Level Outcomes" are achieved through a three-stage RAG (Retrieval-Augmented Generation) process:

1. **Vector Store (ChromaDB):** Hosted on the M3 Max (local storage). It stores the text of the SEC filings you've parsed.
2. **Context Injection:** When you ask about a "Flux," the M3 Max fetches the relevant tables, formats them into a prompt, and sends them over the Tailscale tunnel to the 3080.
3. **Inference:** The 3080 processes the request and returns the analyst summary to your dashboard.

4. Fine-Tuning Roadmap (The 36GB Advantage)

Fine-tuning an 8B model requires ~16GB-24GB of VRAM if using standard methods.

- **The Problem:** Your 3080 (10GB) will likely crash during fine-tuning.
- **The Solution:** Use your **M3 Max** for the fine-tuning phase using **MLX** (Apple's machine learning framework).
- Once the model is fine-tuned on your M3, you can export it as a GGUF file and move it back to the 3080 for high-speed daily inference.

5. Security Protocol

- **No Public IP:** Your desktop remains behind your home firewall.
- **Auth:** Tailscale provides identity-based access; only your logged-in M3 Max can talk to the LLM API.