Home AI Infrastructure Setup Summary

Hardware Inventory

Windows 11 Machine #1 (Primary Ollama Server)

- **GPU:** RTX 3060 12GB
- **Storage:** SATA HDD (NVMe upgrade planned)
- **Purpose:** Primary Ollama inference server
- **Software:** Ollama + Open WebUI
- **Performance:** 70+ tokens/second for 7B models, handles 13B models comfortably

Windows 11 Machine #2 (Image Generation Workstation)

- **GPU:** 2x GTX 1080 Ti 11GB (22GB total VRAM)
- **Storage:** SSD
- **Purpose:** Flux/ComfyUI image generation
- **Secondary Use:** Large Ollama models (30B-70B) when not rendering
- **Advantage:** 22GB VRAM enables complex image workflows and massive language models

MacBook Air M2

- **Specs:** 8GB unified memory
- **Purpose:** Mobile Ollama access while traveling
- **Connection:** Remote access to home Ollama server via Cloudflare Tunnel
- **Performance:** ~40 tokens/second for 7B models (when running locally)

Mac Mini M2

- **Specs:** 8GB unified memory, SSD
- **Purpose:** Plex media server
- **Status:** Running 24/7, low power consumption
- **Optional:** Could dual-purpose as home Ollama API endpoint

Network Architecture

Remote Access Setup

- **Method:** Cloudflare Tunnel
- **Benefits:**
  - Zero Trust security
  - No port forwarding required
  - No exposed IP address
  - Free tier sufficient
  - HTTPS by default

Access Flow

```
MacBook Air (traveling)
   ↓
Cloudflare Tunnel
   ↓
Open WebUI (Win11/3060)
   ↓
Ollama Server
```

Performance Comparisons

RTX 3060 12GB vs Mac Mini M2 8GB

- **RTX 3060:** 70+ t/s for 7B models, 12GB dedicated VRAM
- **Mac Mini M2:** ~40 t/s for 7B models, only ~6GB usable after OS overhead
- **Winner:** RTX 3060 - significantly more usable memory and faster inference

Dual GTX 1080 Ti vs Mac Mini M2 8GB

- **Dual 1080 Ti:** 22GB total VRAM, can run 30B-70B models
- **Mac Mini M2:** Limited to small 7B models maximum
- **Winner:** Dual 1080 Ti by a wide margin - 4x the usable memory capacity

GTX 1040 2GB (Original Setup) vs Mac Mini M2

- **GTX 1040:** Severely limited by 2GB VRAM, SATA HDD bottleneck
- **Mac Mini M2:** 5-10x faster, better for any real LLM work
- **Winner:** Mac Mini M2 decisively

Optimization Strategy

Machine Roles

1. **RTX 3060 Machine:** Dedicated Ollama server for daily 7B-13B models
2. **Dual 1080 Ti Machine:** Primary image generation, secondary large model inference
3. **MacBook Air:** Mobile access, portable computing
4. **Mac Mini:** Always-on services (Plex, potential Ollama endpoint)

Planned Upgrades

- **NVMe upgrade for RTX 3060 machine:** Faster model loading and switching

Why This Setup Works

- Workloads are properly separated (no resource contention)
- Each machine optimized for its specific task
- Remote access enables flexibility without compromising performance
- Redundancy built in (multiple Ollama-capable devices)

Key Takeaways

- **Best for daily Ollama use:** RTX 3060 12GB with Open WebUI
- **Best for large models:** Dual GTX 1080 Ti (when not rendering)
- **Best for mobile:** MacBook Air with Cloudflare Tunnel access
- **Most efficient always-on:** Mac Mini M2
- **NVMe upgrade impact:** Minimal for inference (models stay in VRAM), significant for model switching and large models

Security Considerations

✓ Cloudflare Tunnel provides secure remote access
✓ No exposed ports or IP addresses
✓ Zero Trust architecture
✓ HTTPS encryption by default