

Current State vs Target State

Current State vs Target State

Purpose: Visual comparison of ALAI's architecture today (ANVIL single-point-of-failure) vs the cloud-hosted control plane target state.

Source: `~/system/architecture/cloud-migration-master-plan.md`

TODAY — SINGLE SPOF ARCHITECTURE

ANVIL (makinja-sin-mac-studio)
100.103.49.98

CONTROL PLANE (all-in-one)

Mission Control (mc.js)

└ SQLite mission-control.db
8378 tasks

HiveMind (hivemind.db)

Agent runner (pi-orchestrator)

30 LaunchAgent daemons

Rules/skills/agents (git)

LightRAG (Docker :9621)

Neo4j (Docker :7474/:7687)

Knowledge graph (481MB)

Ollama :11434

qwen3.5:27b (17G)

orchestrator:latest (23G)

alaiml-task/tender/email (3G)

qwen2.5-coder:32b (23G)

bge-m3 + others (~40G)

| LAN only (10.0.0.2)

FORGE (Mac Mini)

Azure swedencentral
4.223.110.181

Supporting services (1 VM)

Standard_B2als_v2, 2vCPU

4GB RAM, 30GB SSD

BookStack (docs)

Vaultwarden (secrets – SPOF)

Planka (boards)

Documenso (signing)

Grafana / Prometheus

Caddy (reverse proxy)

Cost estimate: \$5-53/month

(Azure Founders Hub credit)

Azure Blob (alaibackups0ebb)

system-db-backups

system-git-bundles

bitwarden-exports

Cost: ~\$2.40/month

```
| devstral:24b, qwen2.5-coder |  
| NOT on Tailscale – LAN only |
```

Tailscale mesh: 4 nodes

```
makinja-sin-mac-studio 100.103.49.98  
ab-mac                 100.118.37.71  
basicass-mac-mini     100.104.164.86  
iphone181             100.93.161.73
```

NOTE: ANVIL Ollama :11434 NOT reachable from ab-mac (port timeout verified).

NOTE: 306 files in ~/system/ hardcode localhost:11434 – zero portability today.

SPOF inventory (4 critical):

```
[1] ANVIL dead      → mc.js, HiveMind, agents, LightRAG, Ollama ALL stop  
[2] FORGE dead     → devstral/coder workload stops (Anthropic can substitute)  
[3] Azure VM dead  → Vaultwarden down, secrets inaccessible, agents cannot bootstrap  
[4] Local network  → FORGE permanently isolated (LAN-only, no Tailscale)
```

TARGET — CLOUD-HOSTED CONTROL PLANE + THIN CLIENT

CLIENT (any OS – new laptop, travel machine, etc.)

```
| alai-cli (single installable package) |  
| brew install alai | npm install -g @alai/cli |  
| winget install alai | apt install alai-cli |  
|  
| alai login      → OAuth2 PKCE → Azure AD B2C |  
| alai start      → connects to cloud APIs |  
| alai mc list    → proxies to MC API |  
| alai agent run  → dispatches to agent runner |  
|  
| Claude Code CLI (installed separately) |  
| ~/.claude/ cloned from git on login |
```

```
| HTTPS (Azure Front Door or direct) |  
| Auth: Azure AD B2C JWT |
```

```
|  
| ▼ |  
| CLOUD CONTROL PLANE (Azure Container Apps) |  
| Region: swedencentral (existing subscription) |  
|  
| MC API | Agent Runner API |  
| REST + WebSocket | POST /run |  
| → Postgres | → dispatches agents |
```

HiveMind API
pub/sub
→ Postgres

Skills/Rules Proxy
serves ~/system/
content from Git

Auth API
Azure AD B2C
JWT issuance

Secrets Proxy
→ Bitwarden cloud
(no self-hosted BW)

Azure Database for Postgres (Flexible Server)
Burstable Blms – mission_control + hivemind
(migrated from local SQLite)

Azure Container Registry (private)
MC API, HiveMind, Agent Runner images

┆ Tailscale (encrypted WireGuard)
┆ OR public HTTPS (for Anthropic-only agents)

DATA PLANE (stays on hardware)

ANVIL 100.103.49.98	FORGE 10.0.0.2
Ollama :11434 (primary)	devstral:24b
qwen3.5:27b	qwen2.5-coder:32b
alaiml-task/tender/email	(add to Tailscale)
orchestrator:latest	:11434
LightRAG + Neo4j	(Phase 5)

CLOUD ML FALLBACK (Phase 5)

Together.ai – Llama-3.3-70B \$0.88/M tokens
Triggered only when ANVIL:11434 unreachable

SECRETS (Phase 6 – replaces self-hosted Vaultwarden)

Bitwarden cloud (Teams plan)
\$4/user/month – 1 user = \$4/month
HA by default – Bitwarden's infrastructure
alai-cli integrates via BW CLI at login

Key Differences

Component	Current State (ANVIL SPOF)	Target State (Cloud Control Plane)
-----------	----------------------------	------------------------------------

Mission Control	SQLite on ANVIL disk	Postgres + MC API (Azure Container Apps)
HiveMind	SQLite on ANVIL disk	Postgres + HiveMind API (Azure Container Apps)
Agent Runner	pi-orchestrator on ANVIL only	Cloud agent-runner (Anthropic-powered agents), ANVIL for fine-tuned models
Secrets	Vaultwarden on single Azure VM	Bitwarden cloud (\$4/month, HA by default)
Client Bootstrap	Manual setup, ANVIL-dependent	<code>brew install alai && alai login</code> — under 10 minutes, any OS
Ollama	ANVIL only, FORGE LAN-isolated	ANVIL + FORGE (Tailscale) + Together.ai cloud fallback
Cost	\$27-106/month (mostly hidden by Azure credit)	\$108-165/month (transparent, no hidden dependencies)
ANVIL Offline Impact	Total system outage	Cloud services continue, fine-tuned models pause gracefully

SPOF Elimination

4 SPOFs removed:

1. **ANVIL death** — control plane (MC, HiveMind, agent runner) migrates to cloud. ANVIL offline = Ollama workloads pause, everything else continues.
2. **Vaultwarden VM death** — secrets migrate to Bitwarden cloud (HA by default). No more single-VM secret dependency.
3. **Network isolation** — FORGE joins Tailscale. Cloud services can reach FORGE for code tasks even when ANVIL is down.
4. **Workstation lock-in** — `alai-cli` works from any machine. No more "John only works from ANVIL."

Credit: ALAI, 2026

Revision #2

Created 2026-04-20 16:59:16 UTC by John

Updated 2026-05-31 20:06:14 UTC by John