

Aperture — hoe het werkt + hoe we het veilig inzetten

Kort: Aperture is een AI-gateway van Tailscale. Al je AI-agents (Conductor + bots) sturen hun verkeer door 1 punt. Daardoor zie je centraal: verbruik, kosten per model, wie-deed-wat, audit-logs en quota's. Precies wat je wilde voor kosten-inzicht.

Alles hieronder is gecheckt op de officiële Tailscale-docs (5 juli 2026).

Wat het is

- Een **reverse proxy** die tussen onze agents en Anthropic/OpenAI/Google zit.
- Lost 3 dingen op: API-keys niet meer overal, centraal zicht op gebruik+kosten, en alles auditbaar.
- Herkent elke agent automatisch via de **Tailscale-identiteit** (login/device/tag).
- **Let op: nog beta/alpha** — een experimenteel product, kan bugs hebben.

Werkt het met jouw Max-login?

Ja — via de stand "**passthrough**". Dan:

- Zet je in de agent alleen `ANTHROPIC_BASE_URL` naar Aperture (+ een dummy-key).
- Je zet GEEN echte API-key — de agent stuurt zijn eigen Max-OAuth mee, Aperture geeft die ongewijzigd door.
- De Anthropic-provider in Aperture moet dan op **passthrough** staan.
- Gebruik `http://` (niet `https`) — binnen Tailscale is het toch al versleuteld.

Wat je centraal ziet (observability)

- **Usage/kosten** per gebruiker én per model (bij Max = schatting, geen echte factuur).
- **Model Activity, Tool Use** (je kunt zelfs per tool toestaan/weigeren).

- **Quota's:** uitgaven-limiet per gebruiker/agent/dag ("\$/day") — een echte kosten-rem.
 - **Audit-logs:** zelfs als een beheerder andermans logs bekijkt, wordt dat gelogd.
 - Per bot onderscheiden = geef elke bot een eigen **tag** of eigen node.
-

⚠ Het grootste risico: geen failover

De docs zijn hier heel duidelijk:

- Als de Aperture-node **uitvalt of z'n sleutel verloopt**, dan valt de gekoppelde agent **stil** — er is GEEN automatische terugval naar Anthropic, en een verlopen node is **niet te herstellen** (je moet een nieuwe opzetten).
- Omdat onze agents dan via Aperture "praten", betekent dat: gateway plat = agent offline.

Daarom vóór alles hardening:

1. Maak de `aiforjamal` -node een **tagged device** in Tailscale, én
2. **zet key-expiry uit** voor die node.

Dan kan de gateway niet zomaar stilvallen.

Mijn aanbevolen aanpak (veilig, gefaseerd)

1. **Eerst hardening** — node taggen + key-expiry uit (anders riskeer je de hele fleet).
2. **Provider goed zetten** — Anthropic in Aperture op **passthrough** (jouw Max blijft werken),
`authorization: x-api-key , anthropic_messages: true .Baseurl`
`https://api.anthropic.com`
(LET OP: géén `/v1` erachter — meest gemaakte fout).
3. **Test met Conductor als eerste** — ik zet het erin mét een backup en laat het pas op de volgende herstart ingaan. Verbindt 'ie niet → ik draai het meteen terug (geen schade).
4. **Client-side vangnet** (bouw ik zelf, zit niet in Aperture): als een agent Aperture niet bereikt, val automatisch terug op de directe Anthropic-verbinding. Zo gaat geen bot offline.
5. **Daarna de bots één voor één** — elk met een eigen tag, zodat je per bot ziet wie wat verstookt. Quota per bot als rem.

6. **Behandel het als inzicht-laag, niet als levenslijn** — het is beta + zonder failover.

Altijd een directe uitweg achter de hand.

Praktische snags (nu al gevonden)

- De naam `aiforjamal.tail2d7247.ts.net` resoltet niet vanaf de MacBook (alleen via IP / korte naam). Oplossing: MagicDNS goedzetten, óf de korte naam `http://aiforjamal` gebruiken.
 - Er is ook een **Aperture CLI-launcher** die het koppelen automatisch doet voor Claude Code, Codex, Copilot, OpenCode e.a. — makkelijker dan handmatig env-vars zetten.
-

Wat ik van jou nodig heb

- **Ga ik het gefaseerd opzetten zoals hierboven?** (Conductor eerst, met vangnet + backup.)
- Kun jij in Aperture checken dat de **Anthropic-provider op passthrough** staat?
- En in Tailscale de `aiforjamal` -node **taggen + key-expiry uit** (of ik leg uit hoe).