The Evolution of Agent Hosting

Where your agents live shapes control, cost, data, and capability

cirrius solutions

Where Will Your Agent Live?

Three hosting models. Three very different tradeoffs.

The Three Model:

Self-Hosted

e.g., AWS, GCP, Azure, on-prem

Your infrastructure. Your rules.

* Total control of stack & data
* Bring any model, any framework
* You build retries, audit, scale
* Highest operational burden

Best when: data residency, open-weight models, or scale matter most.

Software-Aligned

e.g., Salesforce Agentforce, Copilot Studio

Lives in your business platform.

* Inherits identity & data model
* Sharing, audit, RBAC for free
* Fastest path to in-app value
* Locked to platform's worldview

Best when: the work already lives in a CRM, ITSM, or back-office platform.

Provider-Hosted

e.g., Anthropic Managed Agents, OpenAI Workspace Agents

You build it. The vendor runs it.

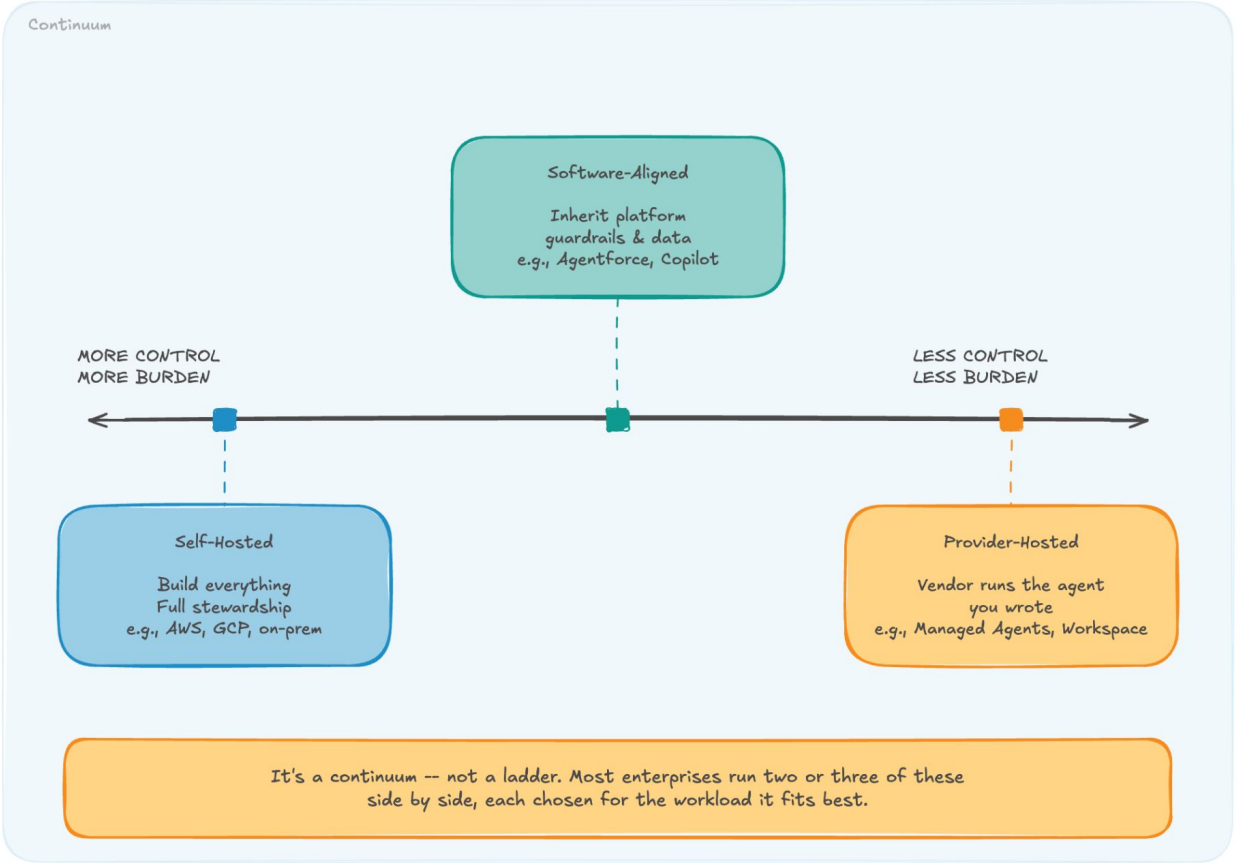
* Vendor owns runtime & scale
* You own logic, prompts, tools
* Sessions & sandbox managed
* Tied to vendor's model family

Best when: you need bespoke agent logic without standing up your own platform.

The choice you make first shapes everything that follows --
your data path, your identity model, your cost curve, and your blast radius.

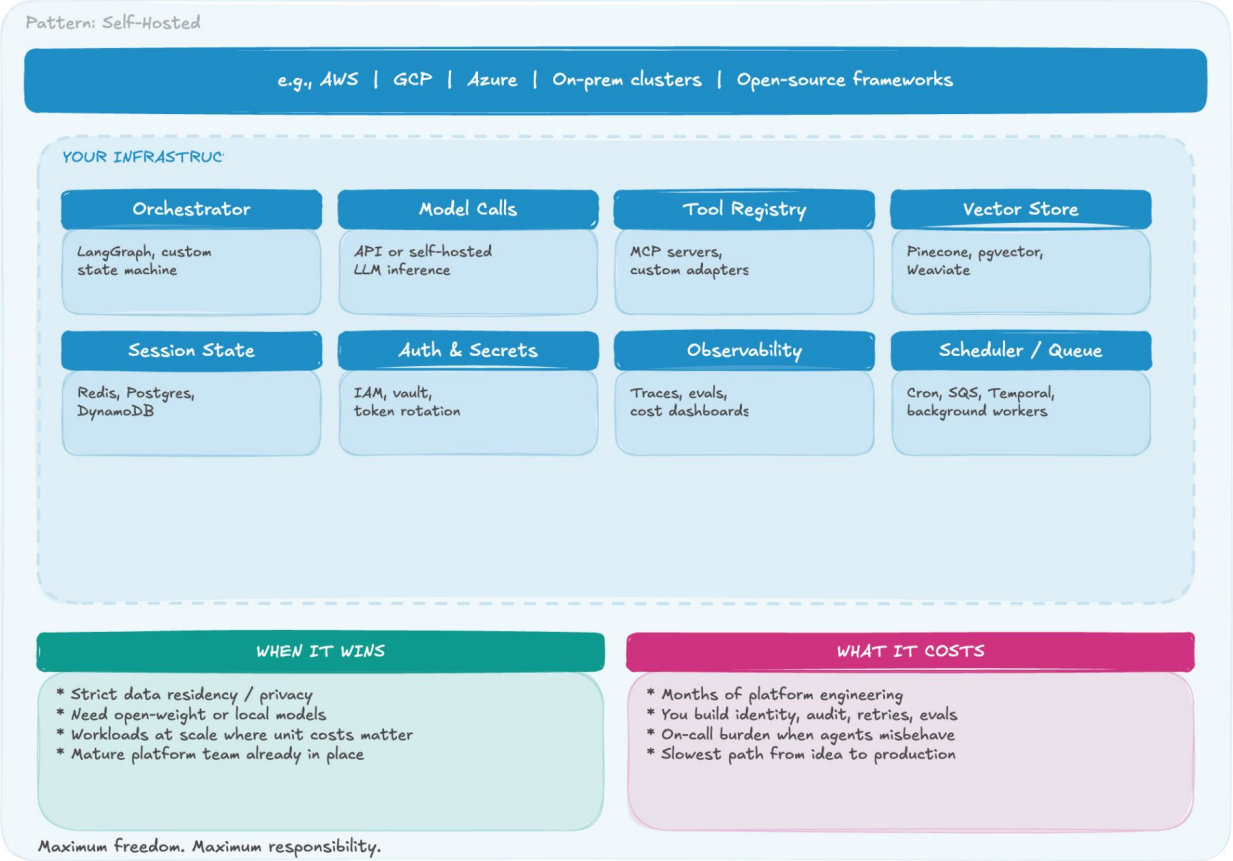
The Hosting Spectrum

Control vs. convenience -- pick your tradeoff.



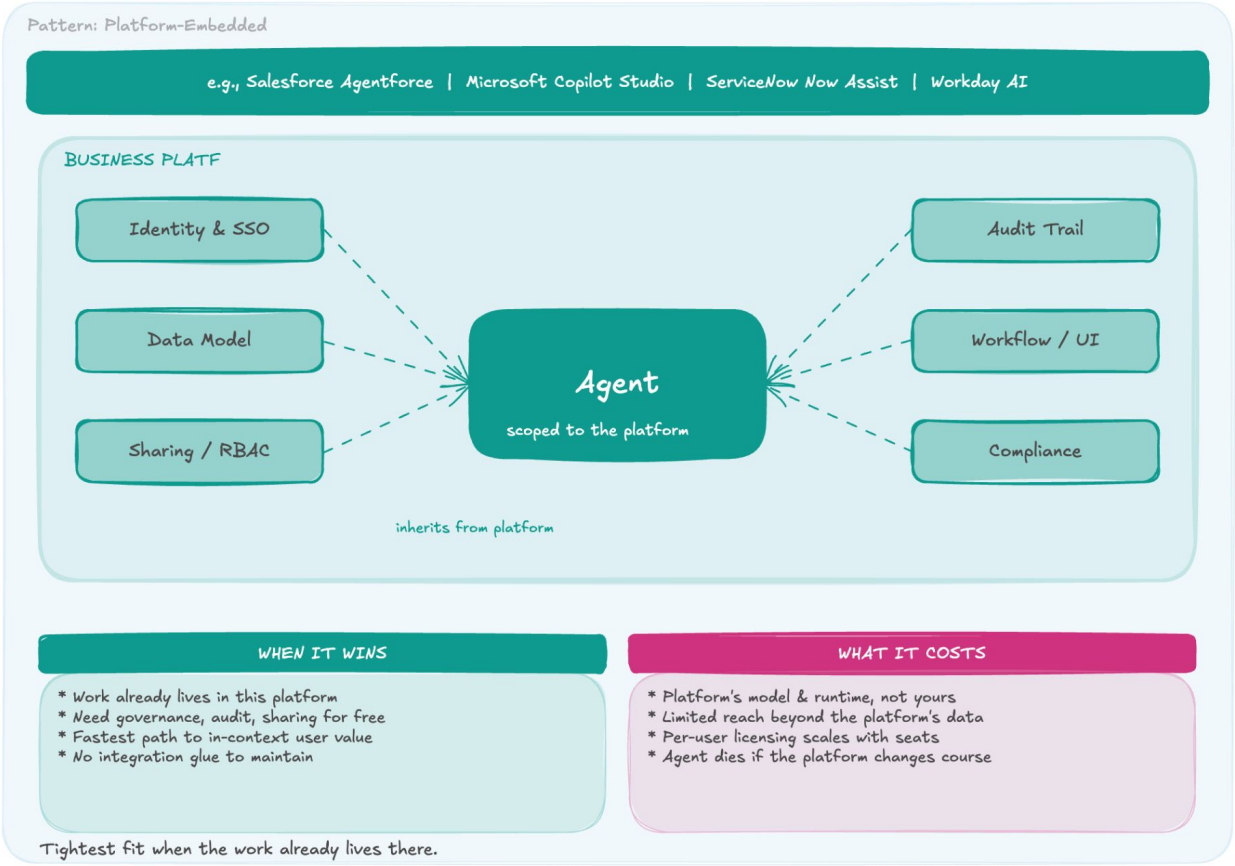
Self-Hosted

You own the entire stack -- from runtime to retries.



Software-Aligned Hosting

Agents embedded in the business platform that already runs the work.

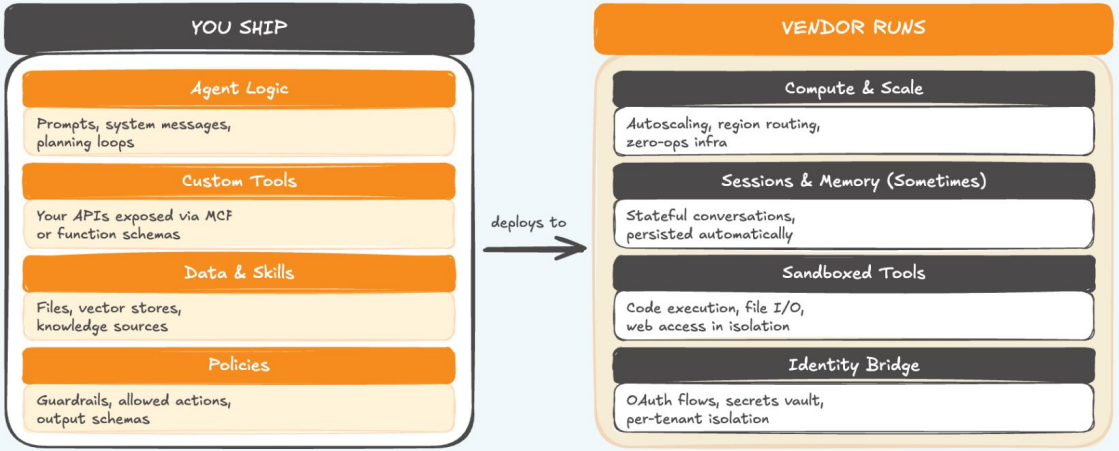


Provider-Hosted

You write the agent. The vendor runs the runtime.

Pattern: Hosted Runtime

e.g., Anthropic Managed Agents | OpenAI Workspace Agents | Bedrock AgentCore | Vertex AI Agents



WHEN IT WINS

* Need bespoke logic, not turnkey
* Want managed sessions & sandboxes
* OK with vendor's model family
* Production maturity in weeks, not quarters

WHAT IT COSTS

* Locked to one model family's roadmap
* Data flows through vendor's cloud
* Per-call or per-seat pricing
* Less visibility into runtime internals

You build the agent. The vendor handles everything else.

How They Compare

Same axes. Very different answers.

Side-by-Side

Dimension	Self-Hosted	Software-Aligned	Provider-Hosted
Operational Burden	Heavy -- you run it all	Light -- platform handles it	Medium -- vendor runs runtime
Cost Model	Infra + ops + tokens	Per-user license (platform seats)	Per token / per call, or per seat
Data Residency	Full control -- any region	Platform's boundary (usually region-pinned)	Vendor's cloud -- check regions
Identity & Access	Build it yourself (IAM, OAuth, vault)	Inherited from the platform	Bring your own auth or use vendor SSO
Time to Value	Slow -- quarters to MVP	Fast -- weeks on platform	Fast -- days to weeks
Vendor Lock-In	Low -- portable agents	High -- bound to platform	Medium-High -- tied to model + runtime
Model Flexibility	Any model, any version	Platform's choice (rarely user-tunable)	Vendor's family (swap in-family only)

Read across each row to spot the tradeoff. No row has a clean winner -- every column trades something to gain something else.

cirrius solutions

How to Choose

Most teams blend two of these. Here is how to pick the first one.

Decision Framework

ASK THESE FIRST

1. Where does the work already live?

If users live in a CRM or ITSM, start software-aligned.

2. How sensitive is the data?

Strict residency or local models => self-hosted.

3. Custom logic, or task-shaped?

Bespoke flows => provider-hosted.
Turnkey work => check vendor-surface flavor first.

4. Do you have a platform team?

No team => avoid self-hosted on day one.

5. How will you observe & evaluate?

If you cannot trace it, you cannot trust it.

COMMON PATTERNS

Platform-First Enterprise

Software-aligned for core CRM/ERP work.
Provider-hosted for bespoke side flows.

Builder-Heavy SaaS

Provider-hosted (embedded) for product features.
Self-hosted for data-sensitive workloads.

Knowledge-Worker Augmentatior

Provider-hosted (vendor-surface) for individuals.
Software-aligned for team-of-record work.

Regulated Industries

Self-hosted for sensitive data paths.
Provider-hosted for non-regulated tasks.

Hosting is not a one-time decision -- it is a portfolio.
Pick the model that fits the workload, expect to run two or three side by side, and revisit as the agent landscape keeps moving.

The right answer is rarely just one of them.

The Big Labs Want to Own the Stack: Harness + Model + Hosting

Two production-grade agent hosting offerings, two very different lenses on how an agent should run

	ANTHROPIC Claude Managed Agents	OPENAI Workspace Agents (in ChatGPT)
	Public beta — April 8, 2026 · Developer / API lens	Successor to Custom GPTs — April 23, 2026 · Business-user / SaaS lens
LENS / FLAVOR	Composable APIs / 'meta-harness'. Separates agent LOGIC from RUNTIME. You bring the brains; Anthropic runs the loop.	Packaged agents living inside ChatGPT. Template-driven, no-code authoring. The agent is a 'co-worker' app, not a service.
TARGET USER	Developers / platform teams shipping production agents (Notion, Rakuten, Sentry, Allianz already deployed)	Knowledge workers + IT admins on ChatGPT Business / Enterprise / Edu. Build via prompt or template, not code.
WHAT'S HOSTED	Orchestration loop · code sandbox · session state · credential vault · persistence · observability · long-context mgmt	Agent runtime · persistent memory · scheduled autonomous runs · approvals workflow · Codex for multi-step execution · in-Slack operation
TOOLS / CONNECTORS	Bring-your-own tools via API + MCP. Open integration surface — anything you can wrap as a tool, the agent can call.	Curated SaaS connector catalog: Slack, Google Drive, Microsoft 365, Salesforce, Notion, Atlassian Rovo (Connector Registry)
PRICING	$0.08 per session-hour + standard Claude API token costs. Usage-based, transparent.	Free during preview through May 6, 2026. Then credit-based on top of ChatGPT seats ($20+/user/mo Business · Enterprise varies).
GOVERNANCE	Platform-managed credentials, sandboxed code exec, debugging + audit trails. Guardrails owned by the developer.	Admin role-based controls · who can build, publish, browse · per-tool restrictions · Compliance API for execution visibility.
KEY DIFFERENTIATOR	Skip months of agent infra. 'Define behavior, we run the loop' — optimized for long-running, multi-step work.	Distribution beats developer experience. Agents land where the work happens — inside ChatGPT, Slack, the SaaS your team uses.
LOCK-IN RISK	Anthropic SDK + proprietary agent format. No open-standard portability today — critics flag harness/runtime lock-in.	Bound to ChatGPT seats + OpenAI runtime. Agent is a tenant inside ChatGPT — moving off-platform means rebuilding from scratch.

TAKEAWAY: Anthropic is selling the runtime to developers — Bring Your Logic, We Run The Loop.
OpenAI is selling the destination to enterprises — The Agent Lives Where Your Team Already Works.
Both are racing to own the full stack: harness + model + hosting. Choose your lock-in deliberately.

cirrius solutions