

MCP Authorization

Enterprise-Readiness Snapshot (May 2025)

Matthew Schwartz

Contents

1	Executive Summary	1
2	Where We Are	1
3	Where the Spec Is Going	1
5	Five-Step Pilot Playbook	2
6	Decision Framework	2
7	Conclusion	3

1 Executive Summary

Enterprises need any new protocol to integrate **securely** with existing identity-provider stacks and to scale **statelessly** behind load-balancers. The **Model Context Protocol (MCP)** is close, yet the ratified spec (v **2025-03-26**) still forces each MCP server to be its **own** authorization server. A draft called **Protected Resource Metadata (PRM)** fixes that by letting servers delegate auth to your IdP. Pilots are viable today; broad rollout should consider waiting until PRM is official.

2 Where We Are

Capability	2025-03-26 Status	Enterprise Impact
OAuth 2.1 + PKCE	YES – standard headers replace ad-hoc auth.	Aligns with zero-trust API gateways.
Dual Role (RS + AS)	NO – server must issue & store tokens.	Statefulness and duplicate audit surface.
Third-party IdP use	WARN – token-mapping (“chaining”) only.	Extra code paths; larger attack surface.
Tool annotations	YES – <code>readOnly</code> <code>destructive</code> <code>authRequired</code> .	Enables blast-radius checks.
Streamable HTTP + batching	YES – one endpoint; fewer round-trips.	Simpler firewall rules.
STDIO transport	NO – no built-in auth.	Local plug-ins trust caller implicitly.

3 Where the Spec Is Going

Draft Change	Benefit	Key Link
Protected Resource Metadata (PRM)	Decouples auth; servers stay stateless & point to an external AS.	PR #284
RFC 9728	PRM now an IETF spec — tooling can rely on stable semantics.	RFC 9728
Updated Auth Spec (Apr 2025)	First MCP doc to <i>require</i> PRM (backed by Anthropic / Microsoft / Auth0).	den.dev
AWS Deployment Guidance	“MCP Auth Service” fronts servers with Cognito + WAF.	AWS guide

Area	Status	What to Do Now
Auth & tokens	● Green (with PRM)	Terminate TLS + JWT validation at the gateway; keep servers stateless.
Token lifecycle	● Yellow	Use short-lived self-encoded JWTs; avoid DB-backed reference tokens.
Least privilege	● Yellow	Publish a scope registry (<code>tools.read</code> , <code>tools.write</code> , <code>admin</code>).
Observability	● Yellow	Inject trace IDs at the gateway; request JSON audit feeds from vendors.
Prompt / tool safety	● Red	Add out-of-band LLM safety filters until the spec adds hooks.

5 Five-Step Pilot Playbook

1. **Run PRM branches only** → point them at your IdP (Azure AD, Okta, Cognito).
2. **Front every MCP endpoint with API Gateway + WAF** → rate-limit, log, validate JWTs.
3. **Publish enterprise scopes early** → stop scope sprawl before it starts.
4. **Wrap STDIO plug-ins with a local proxy** injecting signed user tokens.
5. **Threat-model third-party servers** → SBOM, pen-test, supply-chain scan.

6 Decision Framework

Question	If “Yes”	If “No”
Run draft PRM in a ring-fenced env?	Pilot now.	Re-evaluate after next spec.

Question	If “Yes”	If “No”
Can gateway enforce JWT & scopes?	Servers stay stateless .	Budget for token-store ops.
Ready to own a scope catalogue?	Achieve least-privilege .	Risk over-privileged defaults.

7 Conclusion

MCP’s March release shipped OAuth 2.1 but made every server its own auth system. The April **PRM** draft fixes that flaw, letting organisations integrate MCP with existing IdPs **without** extra state or duplicate audit surfaces.

Bottom line

* **Controlled pilots** — yes, if you run PRM branches behind standard cloud controls.

* **Production at scale** — wait for the next spec that embeds PRM, standardises scopes, and defines audit hooks.

References & Key Links

- [MCP Spec \(2025-03-26\)](#)
- [PRM Pull Request #284](#)
- [RFC 9728 — Protected Resource Metadata](#)
- [Christian Posta blog](#)
- [Den Delimarsky article](#)
- [AWS Guidance](#)