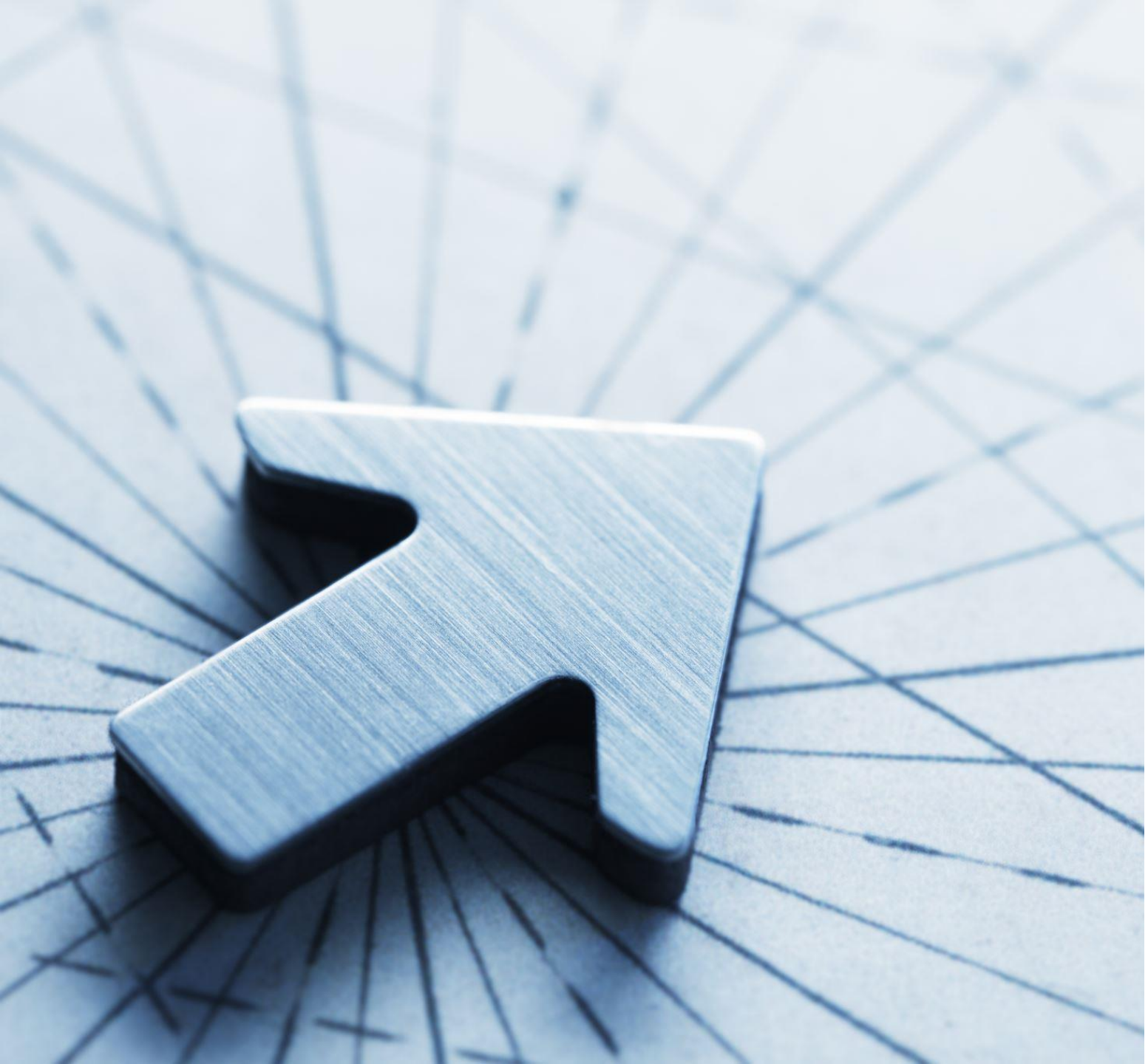


From the lab:

Improving AI Identity, Trust, & Provenance using OpenID Federation

Chris Phillips – Chris@adiuco.com

About Chris

- Independent identity architect
- Serial collaborator in federated id, standards, DevSecOps
- 10+ yrs operating Research and Education multi-lateral federation, SAML & roaming RADIUS SSO infrastructure
- Contributor to IdPro.org exam, Certified IdPro
- Co-founder VeriMe.coop

Chris Phillips, CIDPRO

Identity & AI Architect | Federation, Strategy, Execution

Outcomes for today

- Expand world view and assist the dialogue on
  - Describing the challenges more clearly
  - What OpenID Federation is and how it can assist
  - Ways to pass on our hard earned lessons in this new medium

Challenges

- **What we see**
  - Shadow MCPs & unmanaged APIs
  - Reinvented allow lists, brittle configs
  - Inconsistent onboarding & drift
  - Unknown provenance of endpoints

- **Why it matters**
  - Data exfiltration / prompt injection
  - Rug-pull MCPs, impersonation
  - Supply-chain compromise (unsigned images)
  - Compliance gaps & audit failure

- **Why it repeats**
  - No multilateral trust fabric
  - Clients skip pre-flight validation
  - Bilateral sprawl scales poorly
  - Governance signals aren’t portable

Federation guides whom to trust.

OAuth/OIDC still decides what you can do.

Corollory (provocative?):

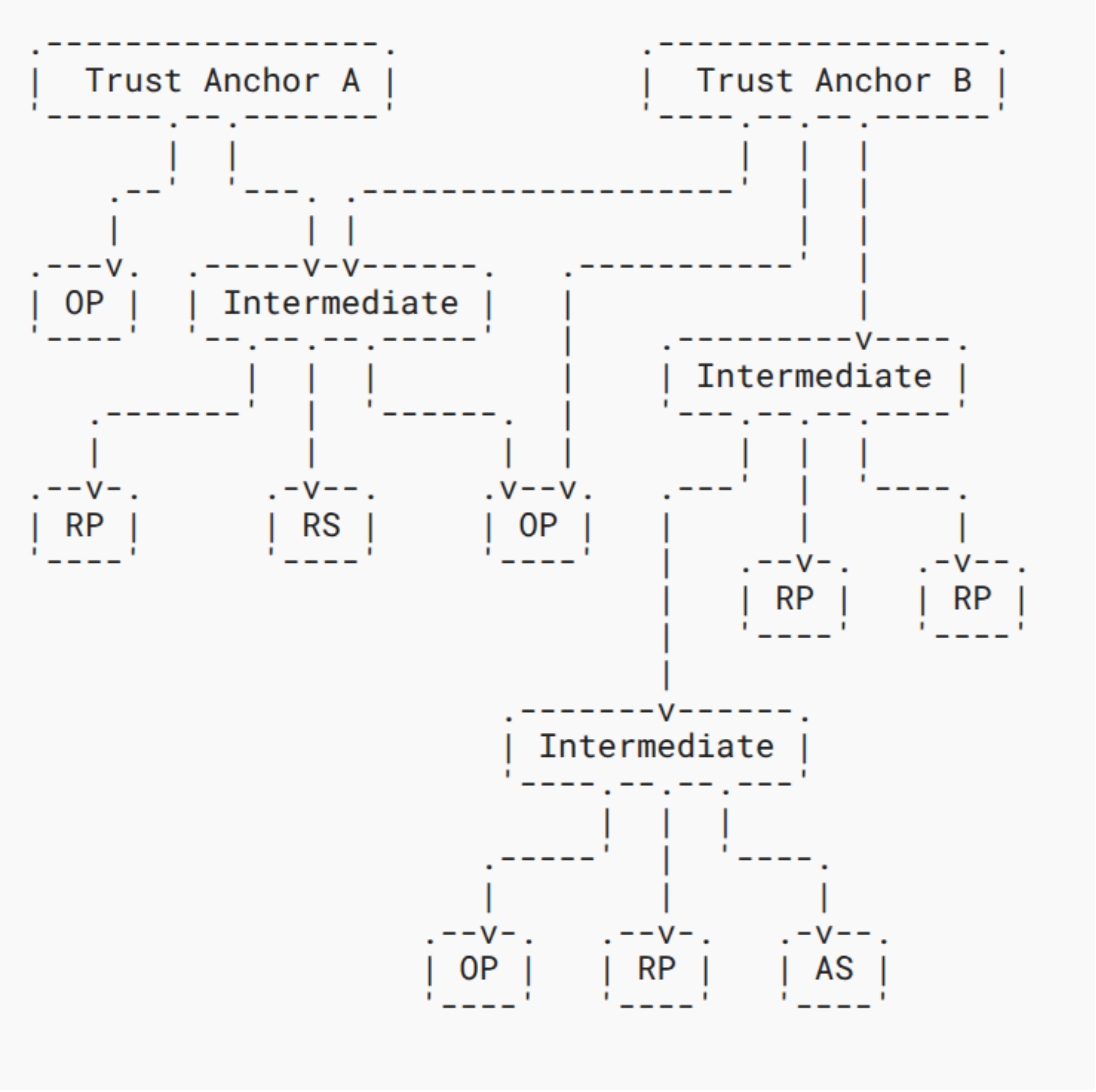
MCP Registries assist calculus on trust but not comprehensive.

Observations: Not portable across protocols, has runtime obligations to scale

My take: Registries & OpenID Fed complement & could amplify each other...

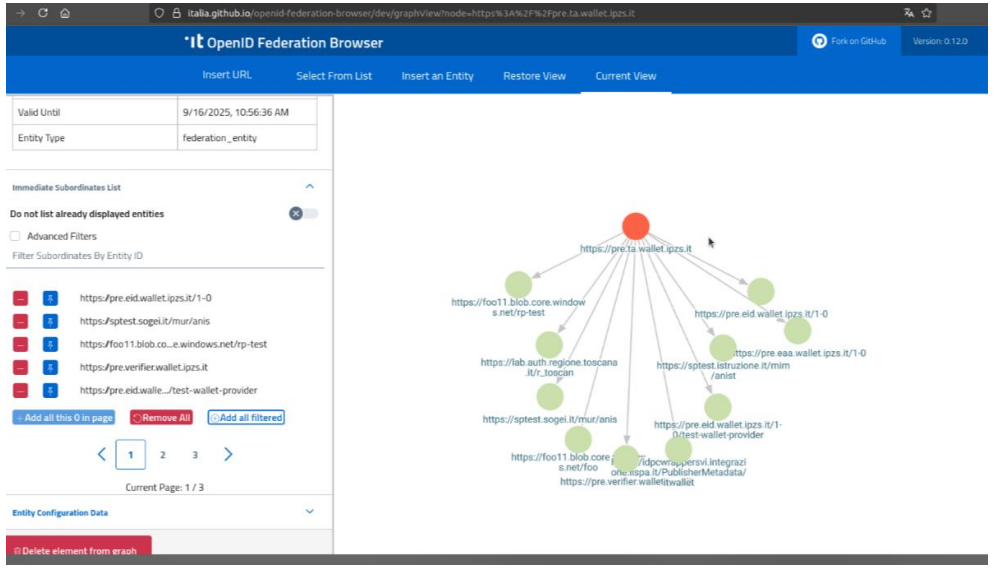
OpenID Federation

- **The trust fabric**
  - Each entity (OP/RP/RS) exposes an Entity Configuration
  - Receives signed Entity Statements, a trust chain up to a Trust Anchor (a CA).
  - Separate process from runtime use
- **Trust marks signal membership / inclusion**
  - Appear in signed JWT attestations
  - signal membership in / conformance to a set of policies or evidence of action
  - Portable across domains.
- **Verifying & Deciding (pre-flight)**
  - Anyone verifying trust resolves the chain, validates marks under the anchor, and enforces policy-as-code ALLOW/DENY.
- **Transacting (runtime)**
  - If allowed, run standard OAuth2/OIDC (Auth Code + PKCE, strict aud).
  - Tokens remain per-recipient; no mark = no connect. (TBD)
  - JWTs consumable with existing OIDC / OAuth libraries infra



OpenID Federation in the field..

- Italy's SPID system (Public Digital Identity System)
- OpenID for Verifiable Presentations for wallets management and presentment of creds
  - Formats: W3C Verifiable Credentials Data Model, ISO mdoc, and IETF SD-JWT VC
- eduGAIN.org OIDFed pilot - R&E's (research & education) 10,000 entity SAML2 fed



- Protocol is mature enough to deploy at nation level services
- implementations at various maturity levels

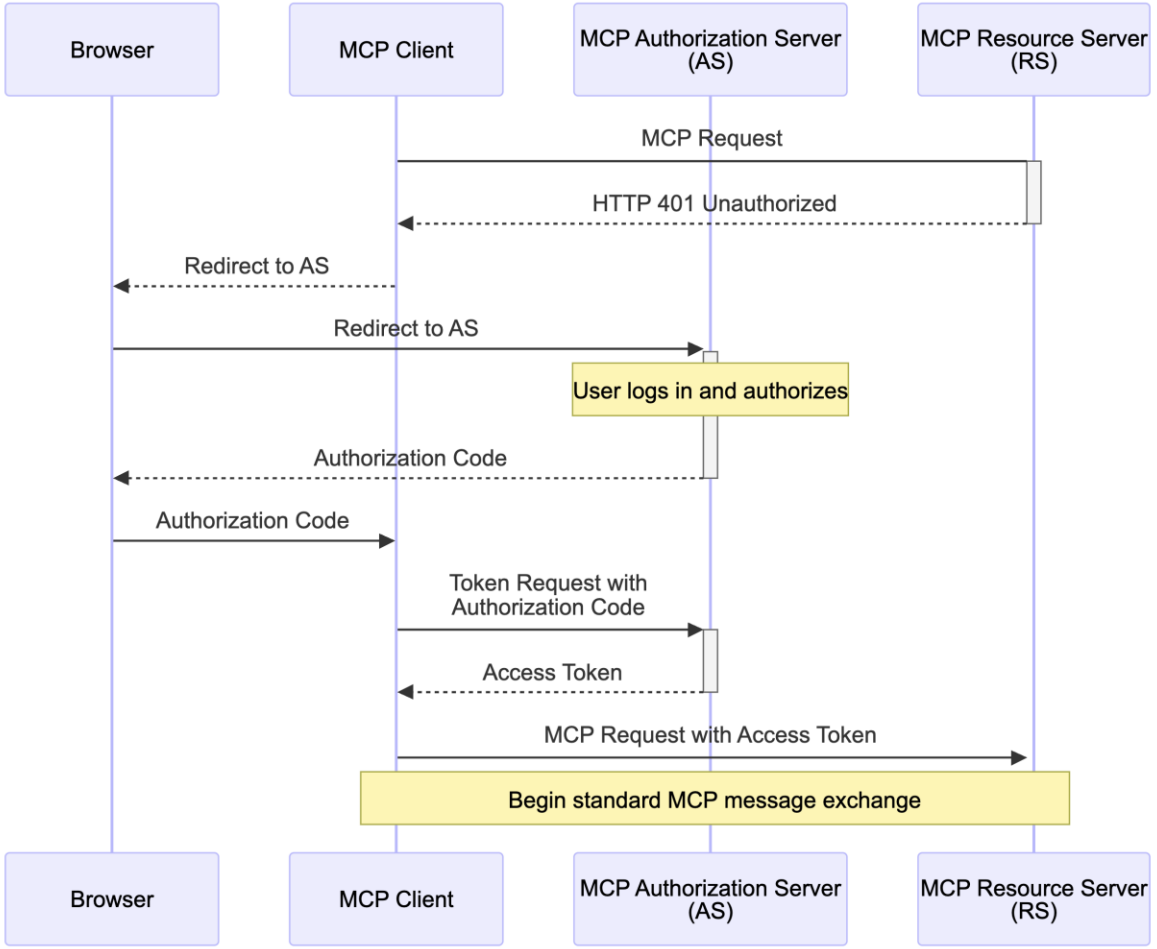
Candidate use cases

- **Rogue MCP defense**
  - Only connect to MCP servers bearing a ‘Trusted MCP Server’ mark from your anchor
- **DevSecOps/SBOM attestation**
  - Require a trust mark asserting the server runs a cosign-signed image at a known digest
- **Function-level authorization input**
  - Use marks/claims to scope which MCP tools/functions a principal may invoke (dev, prod etc)
- **Federation-of-one**
  - Local trust anchor for single-user/maker setups; same mechanics, smaller blast radius
- **License marks**
  - Trust mark for your customers to know which components you bless
  - Can instances of functionality be licensed? (e.g. activation key delivery?)
- **Crypto agility**
  - Rotate federation keys quickly; adopt PQC when ready without breaking runtimes

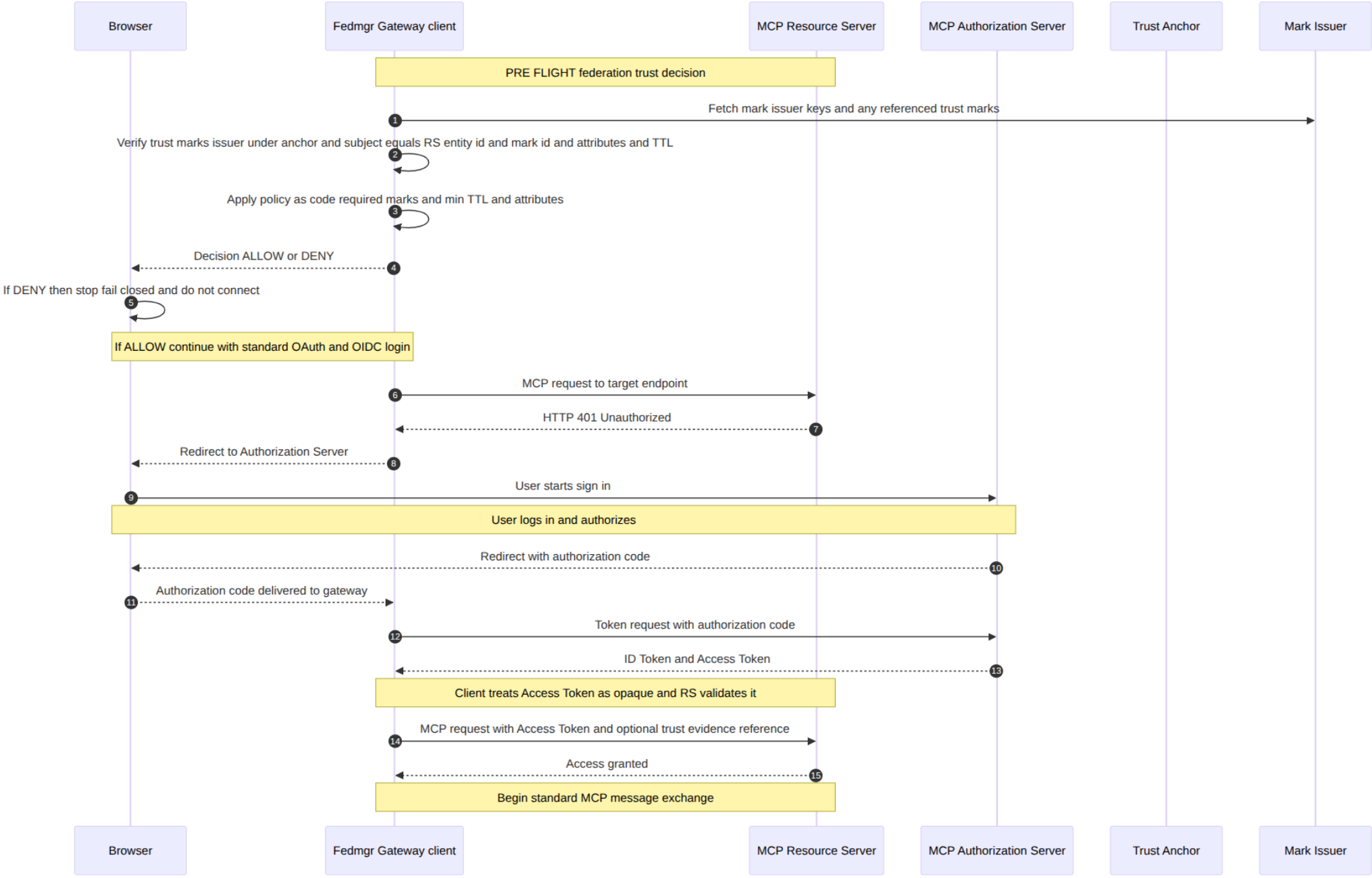
Trust marks are evidence of what was done to *earn* its assignment.

Works in both directions, clients should 'fail secure' by not connecting by default if trust mark doesn't exist.

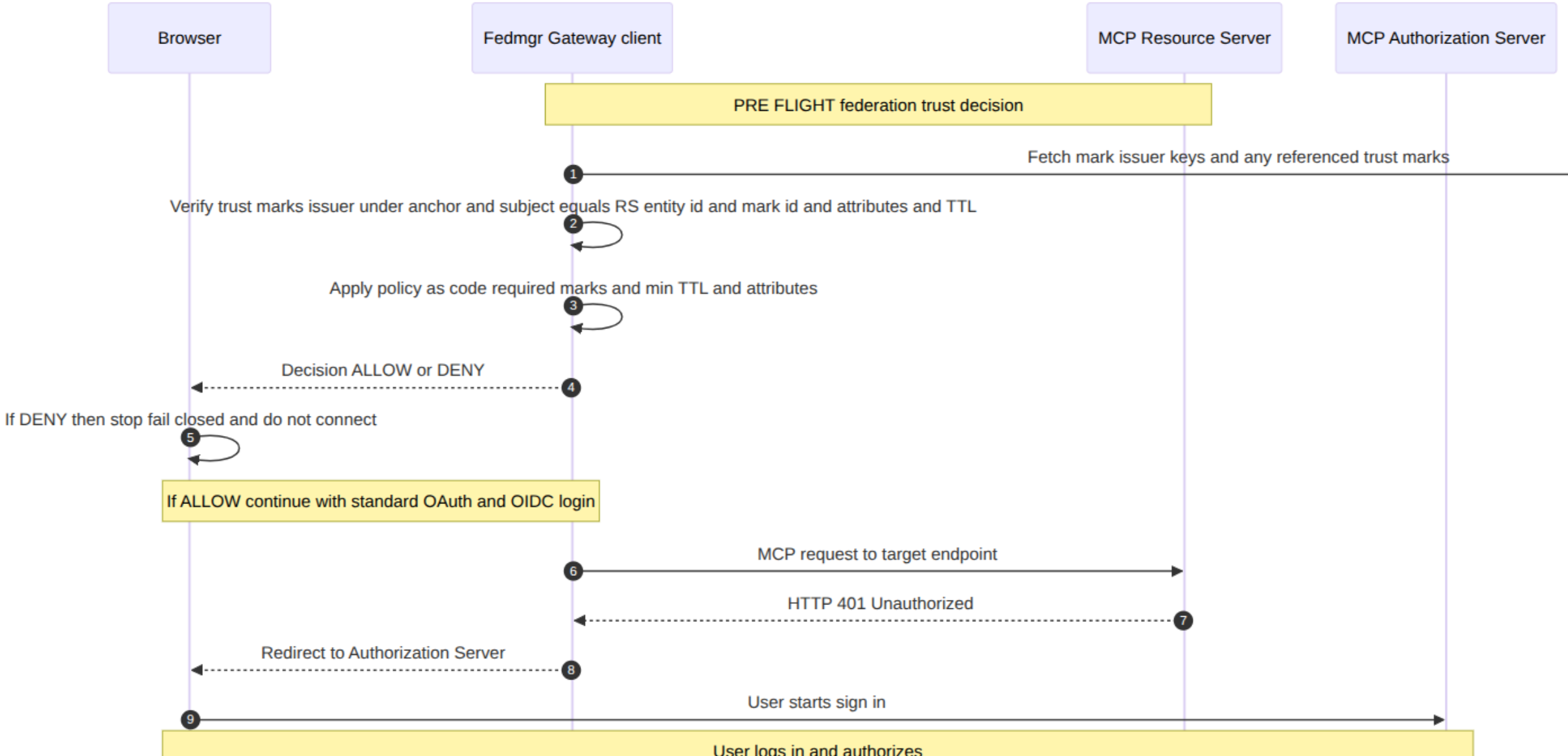
Existing MCP flow



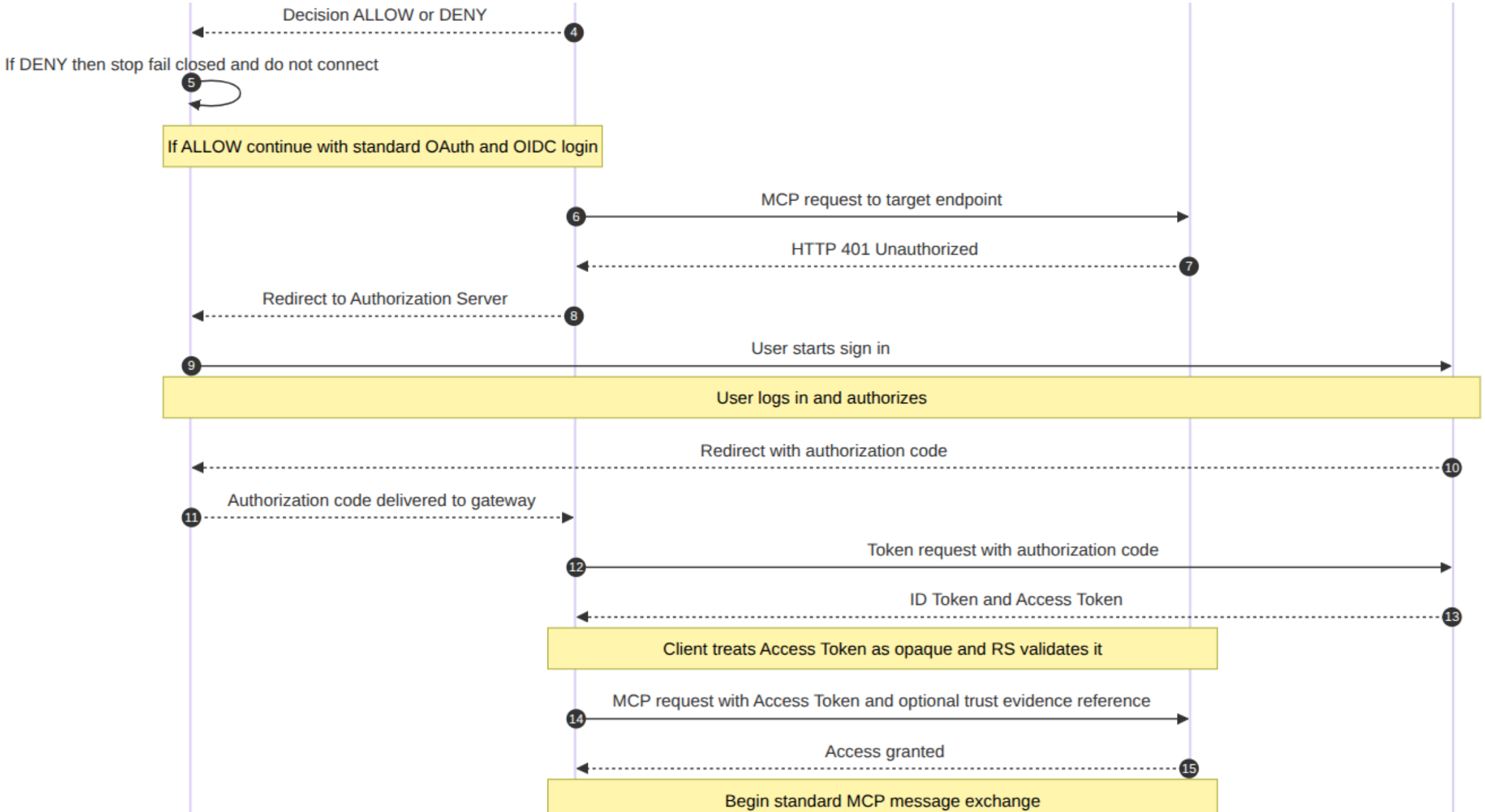
OpenID Federation MCP flow



OpenID Federation MCP flow



OpenID Federation MCP flow



Deployment patterns

- **Audience of 1 endpoint to MCP GW**
  - Sign into github
  - a trust anchor exists and has enrolled 1 MCP entity + the gateway
    - GW re-mints the JWT with trust marks
  - Use the JWT in VSCode
  - Good for proxy of other MCPs
- **Gateway with token exchange & re-audiencing**
  - Same as above but add..
  - RFC 8693 token exchange; re-audience to RP of new audience.
  - fedmgr validates subject_token and client auth, applies policy.
  - It mints a new ID-style assertion targeted to RP-Cross and reattaches fed_marks[].

Where next?

- **Gathering use cases to refine technical elements**
  - Have one? Reach out!
- **Refine Deployment Models:**
  - **Audience-of-1 –** Gateway endpoint with 1 single MCP proxying others (MetaMCP?)
  - **Token Exchange -** Gateway canonical audience vs distributed re-audience (RFC 8693).
- **Release v 0.x of NPM packages that have a demo federation**
  - Improved demo abstracting out functions to library for managed federation
  - boilerplate MCP elements
- **Infrastructure-wise**
  - Opportunity to offer submission to Anthropic as SEP?

References

- https://simpleidserver.com/docs/tutorial/openidfederation
- https://openid.net/specs/openid-4-verifiable-presentations-1_0.html
- https://docs.italia.it/italia/spid/spid-cie-oidc-docs/it/versione-corrente/la_federazione_delle_identita.html
- https://openid.github.io/OpenID4VP/openid-4-verifiable-presentations-wg-draft.html#section-11.2
- https://events.geant.org/event/1946/
- https://wiki.geant.org/spaces/eduGAIN/pages/1072398451/eduGAIN+-+Open+ID+Federation+Pilot
- https://github.com/GEANT/edugain-oidf-pilot

Thank you!

- Questions?
- Use cases to share?
- Looking for deeper engagement?
- Email: Chris@adiuco.com