OAuth 2.0 Scope Aggregation for Multi-Step AI Agent Workflows

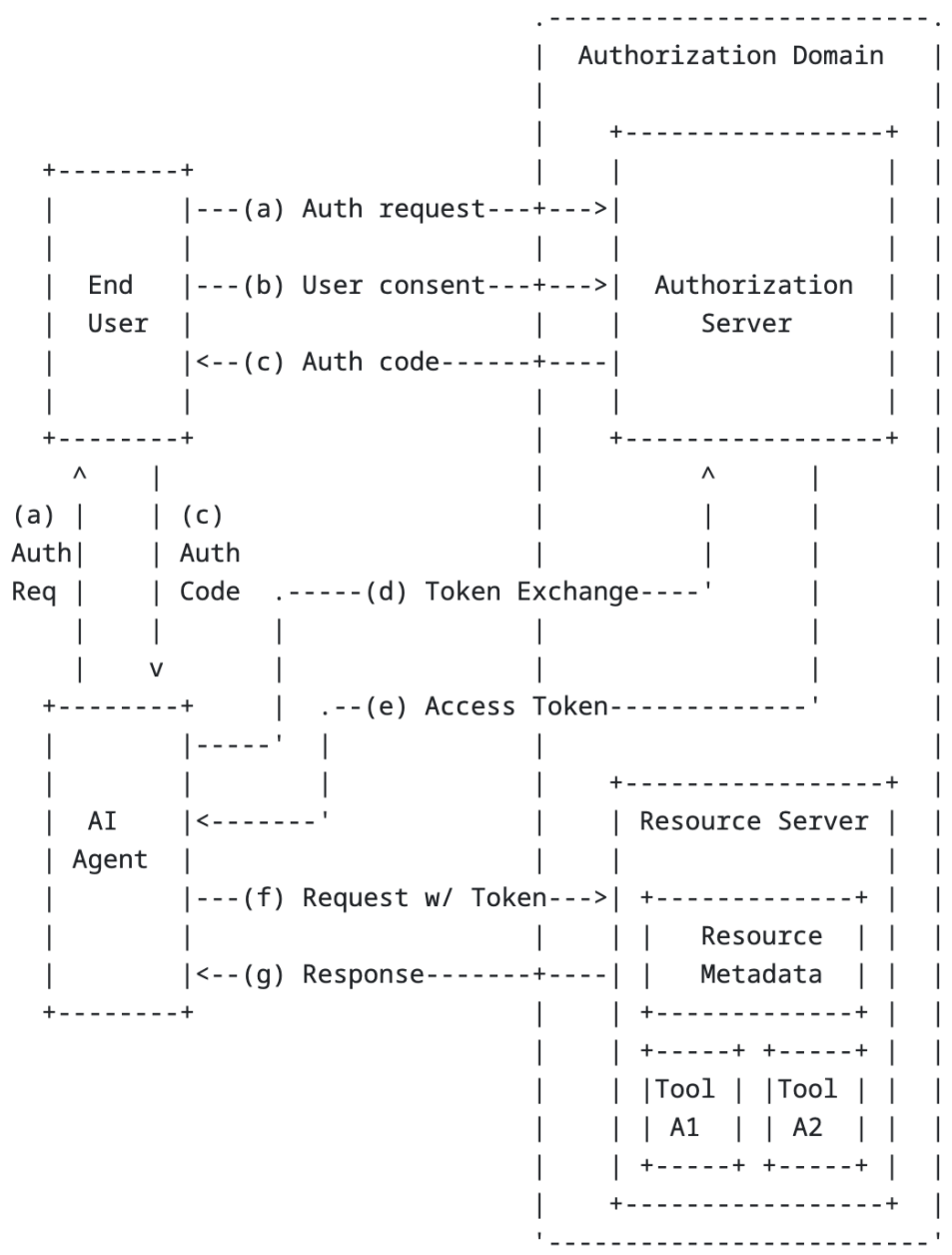
https://datatracker.ietf.org/doc/draft-jia-oauth-scope-aggregation/

Presenter: Yukuan Jia

jiayukuan@huawei.com

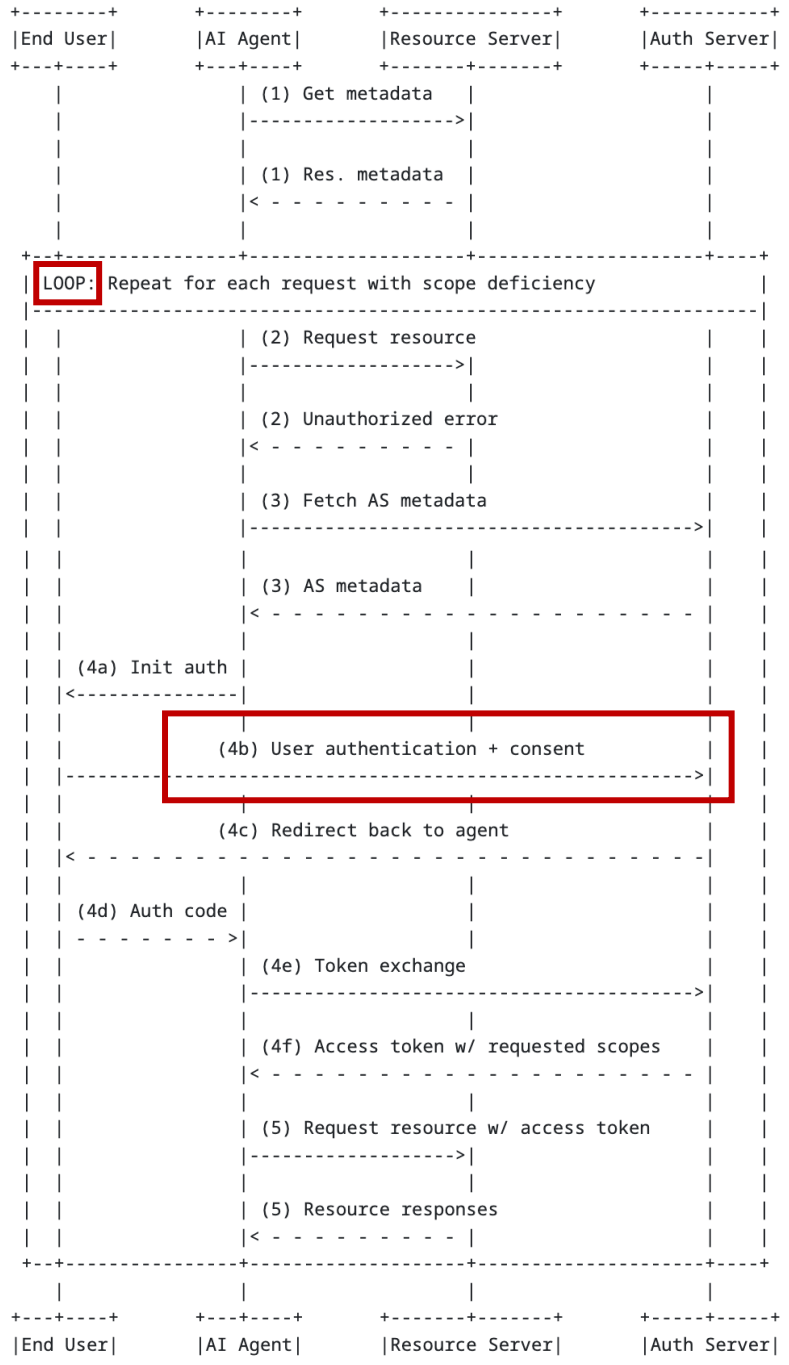
Background

- Advances in LLM enables AI agents to plan and execute multi-step workflow to accomplish tasks
- With agent communication protocols, an AI agent can
  - call third-party tools (e.g., via MCP)
  - delegate sub-tasks to other AI agents (e.g., via A2A)
- These interoperations require authorization, usually by OAuth 2.0
  - AI agent - OAuth client
  - End user - Resource owner



Problem

- Existing agent protocols (e.g. MCP) adopt OAuth 2.0 AuthZ flow in a reactive and challenge-triggered way
  - OAuth 2.0 Protected Resource Metadata (RFC9728)
  - OAuth 2.0 Authorization Server Metadata (RFC8414)
- Scenario: An agentic workflow spans across multiple OAuth protected resources with
  - **distinct RS**: e.g. Maps / Calendar / … MCP services
  - **distinct scopes**: e.g. read / write / admin
- Problems:
  - User is required to respond to multiple consent prompts, causing user friction and fatigue
  - Missing a single AuthZ requirement -> mid-workflow failure

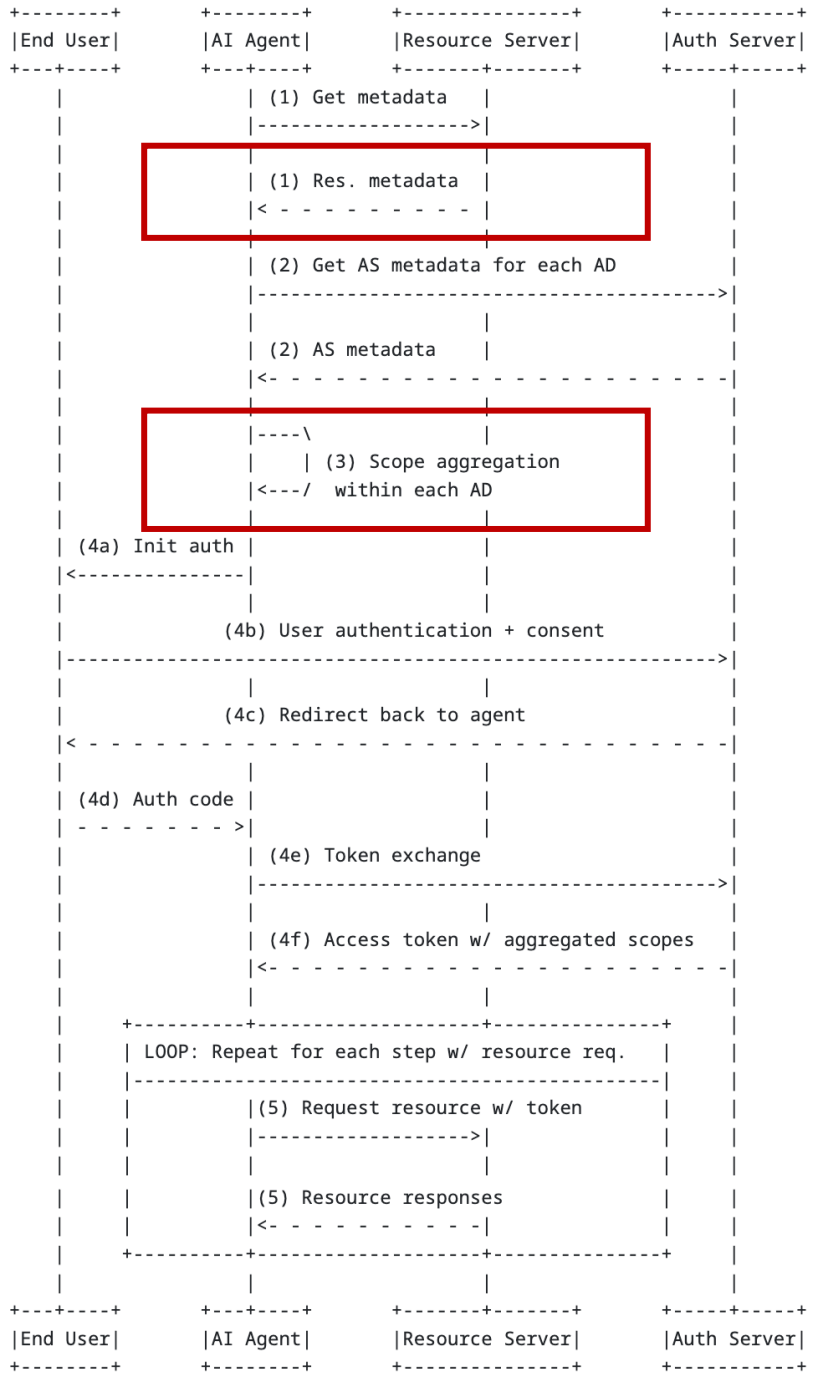


Solution

- Note the security requirement in the resource metadata

```
{
  "name": "resource_identifier",
  "description": "...",
  "input_schema": { ... },
  "security": {
    "type": ["oauth2"],
    "scopes": ["scope_A"],
    "as_metadata": "https://server.example.com/.well-known/..."
  }
}
```

- Aggregate the required scopes within a single authorization domain (authorizable by a single trust anchor)
  - Example A: Step 1 requires ‘mail.read’, Step 2 requires ‘calendar.write’ Aggregated: [‘mail.read’, ‘calendar.write’]
  - Example B: Step 1 requires ‘mail.read’, Step 2 requires ‘mail.admin’ Aggregated: [‘mail.admin’]
- Initiate an OAuth 2.0 AuthZ flow for the aggregated scopes



Summary and Future Work

- Proposes to aggregate the scopes across an agent workflow and initiate a single AuthZ flow, leveraging the security member in the resource metadata

- Future work: extend ‘scope’ to more fine-grained ops?
  - Intent AuthZ (whitelisted operations and constraints)

- Suggestions / Comments?