

AI-First Ops: Streamlining Cisco Infrastructure with Agents, Protocols, and Automation

cisco Live !

Shannon McFarland
VP, Head of Cisco DevNet and Cisco OSPO
CCIE #5245

Cisco Webex App

Questions?

Use Cisco Webex App to chat with the speaker after the session

How

- 1 Find this session in the Cisco Live Mobile App
- 2 Click “Join the Discussion”
- 3 Install the Webex App or go directly to the Webex space
- 4 Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until 14 November 2025.



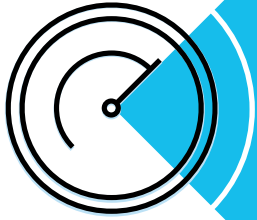
<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKCLD-2070>

Agenda

- 01 Why AI-first Ops Now?
- 02 Operational Challenges
- 03 What does AI-first Ops Mean?
- 04 New tools, protocols, and projects
- 05 Use Cases
- 06 Summary

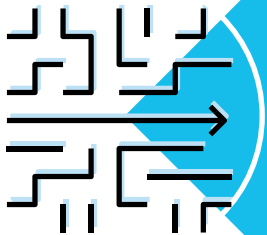
Why AI-first Ops Now?

Increasing Complexity Demands a New Operational Paradigm



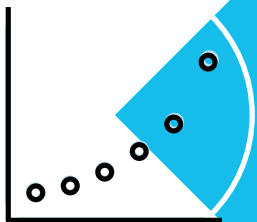
Speed Up

- Design: Streamline patterns, reuse templates
- Deployment: CI/CD, IaC Pipelines
- Feedback: Shorter iterations & validation cycles



Automate

- Test: Unit, integration, regression
- Documentation: Code-driven, auto-generated
- Deployment: Pipeline-driven, self-service



Quality Improvement

- Security: Shift-left checks, scanning
- Observability: MELT++
- Accountability: Ownership, collaboration, governance

What does AI-first Ops Mean?



AI-first Ops integrates AI to automate and transform IT/business operations, enhancing decision-making, automating tasks, and enabling proactive issue prevention without major infrastructure changes. It supports gradual AI adoption to drive operational excellence and innovation

Cisco Data Intelligence CVD

AIOps and AI-first Ops

Add clarification about AgenticOps

- AIOps:

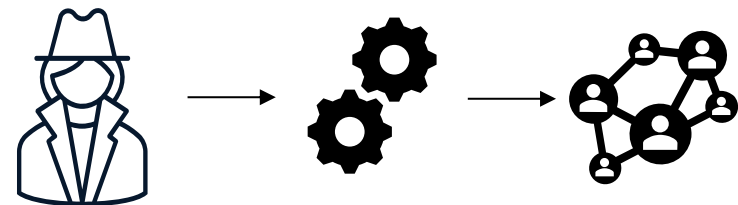
- Additive and often focused on data analysis, anomaly detection, etc.
- Using AI to improve the operations of a non-AI/AI environment

- AI-first Ops:

- AI-first strategy – Think API-first
- AI native from the start
- Focused on agentic systems, proactive orchestration, and integrating AI at every layer



**AI helping Humans Observe,
Detect, and Analyze Operations**



AI Embedded into the Operational Process

AI-first Ops

**Data Collection, RCA,
Anomaly detection,
Enrichment**

– OTel, RAG, MCP, Fine-tuning



Data



AI Augmenting Human Expertise

– Playbook, Runbook, script generation/refactoring

LLM/SLMs

AI Ops



Agents Orchestrating Work

– IaC, API, Tools

AI Agents



Communications

**Communications, Collaborations,
Notifications**

– HITL, Per-component
communications, event notifications

Operational Challenges

The Modern Operator & Developer

Understand Requirements

- Ever-increasing complexity
- Ever-increasing pace

Design

- Tool & Tech Sprawl – APIs, Microservices, etc.

Diverse Collaboration

- Tool Sprawl
- Forced Interaction



© BalanceFormCreative / stock.adobe.com

Write / Debug / Test Code

- Languages
- Frameworks
- Tools

Support

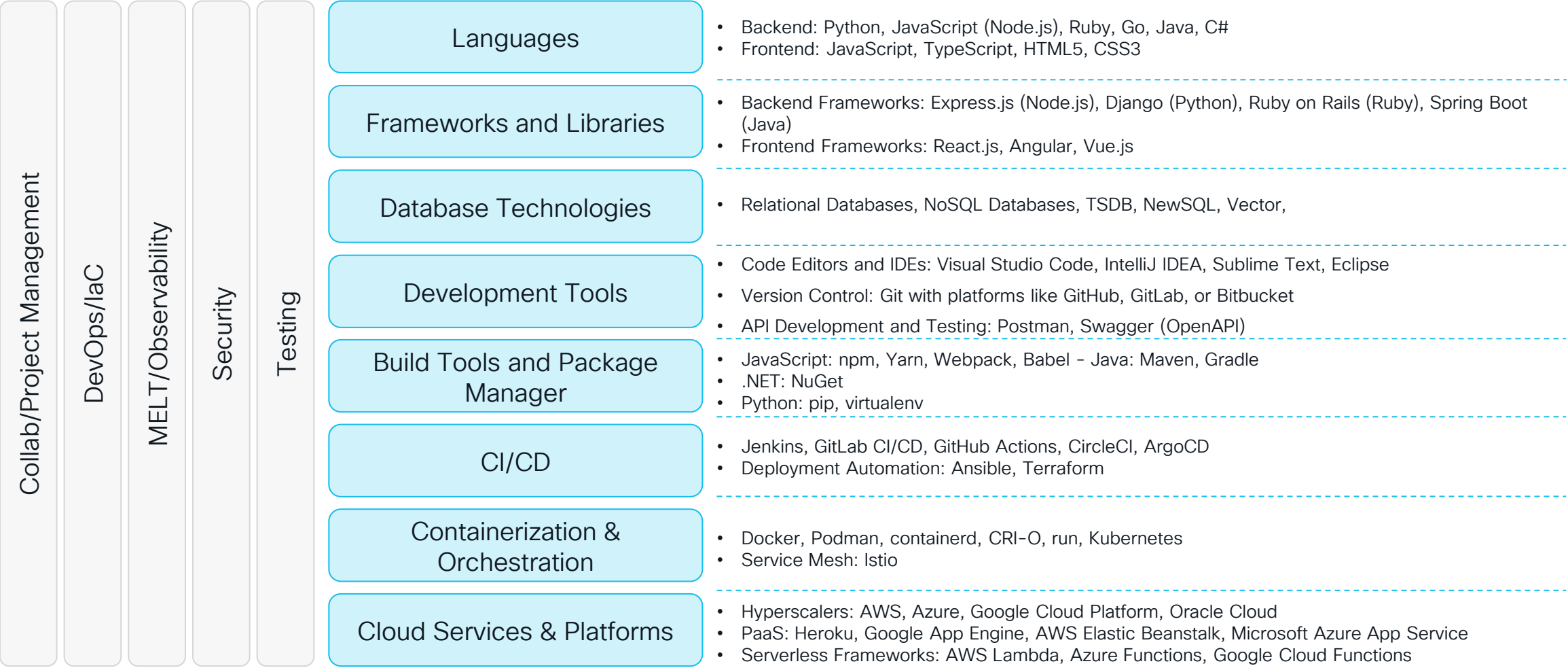
- SRE
- Service Sprawl
- Collaboration Sprawl

Documentation

- APIs
- IaC
- Code

Modern Operator & Developer Stack

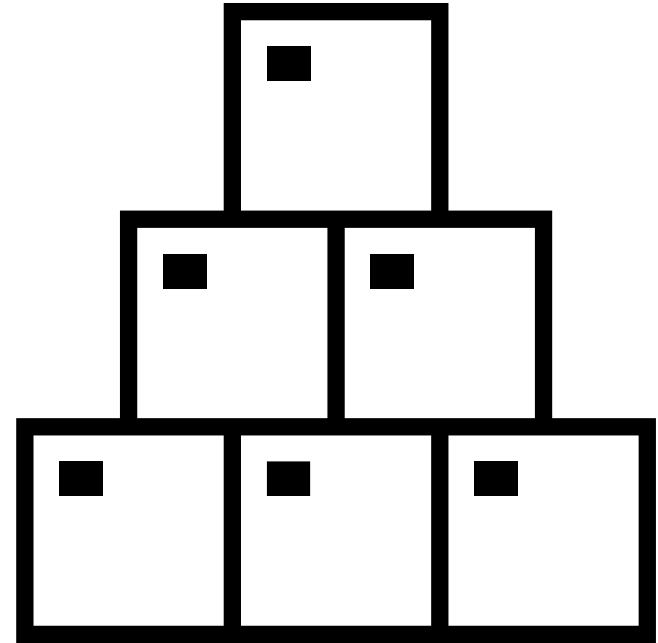
Tool/Service View



New Tools, Protocols, and Projects

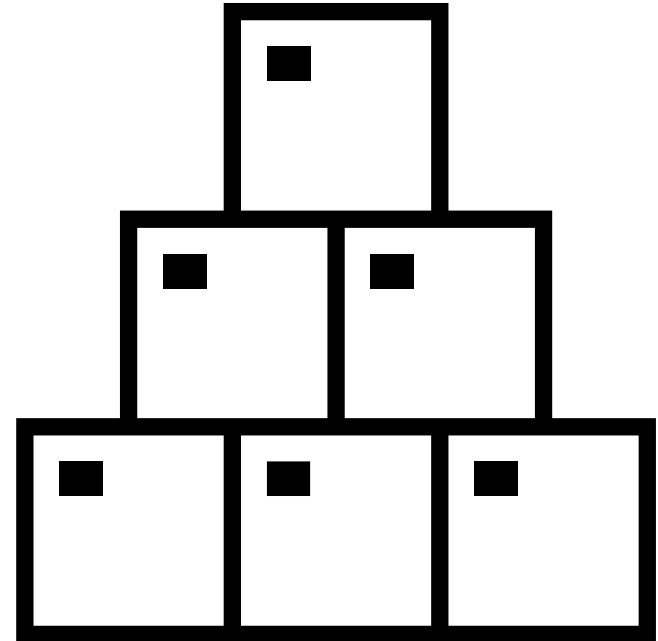
Critical Components of an AI-first Ops Deployment

- Model Context Protocol: AI-driven interaction with existing software services
- AI Agents
- RAG systems
- VectorDBs
- Existing software services:
 - Standard application programming for deterministic outcomes
 - Infrastructure-as-Code
 - CI/CD



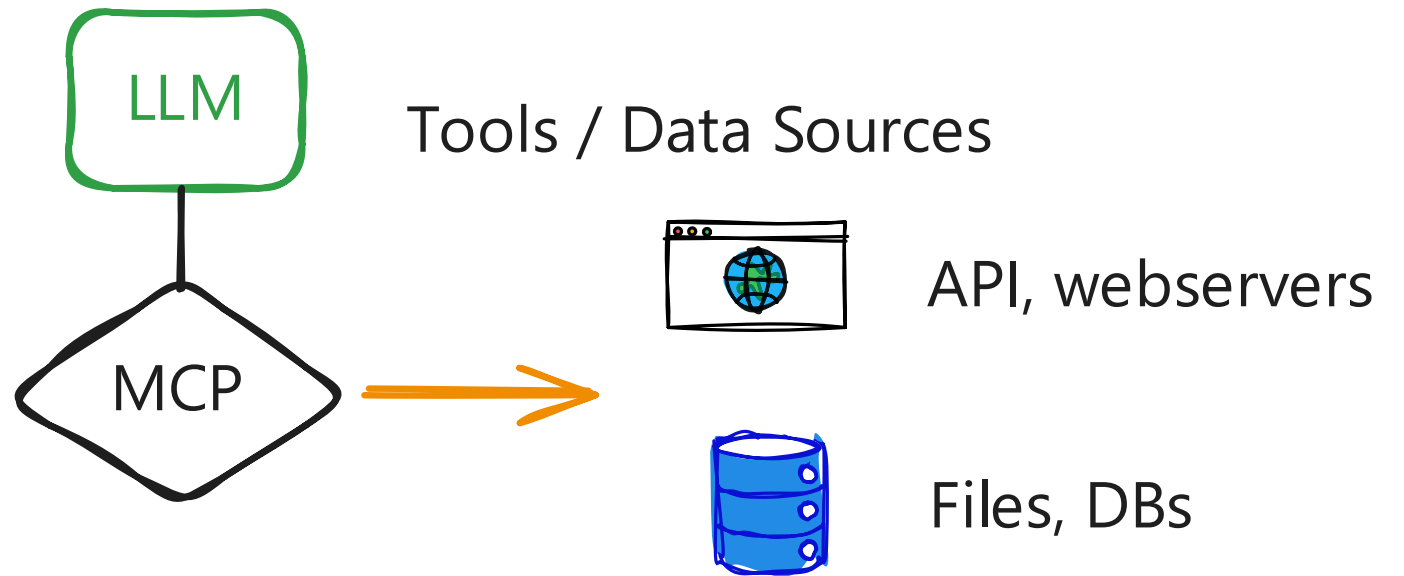
Critical Components of an AI-first Ops Deployment

- **Model Context Protocol**: AI-driven interaction with existing software services
- **AI Agents**
- RAG systems
- VectorDBs
- Existing software services:
 - Standard application programming for deterministic outcomes
 - Infrastructure-as-Code
 - CI/CD



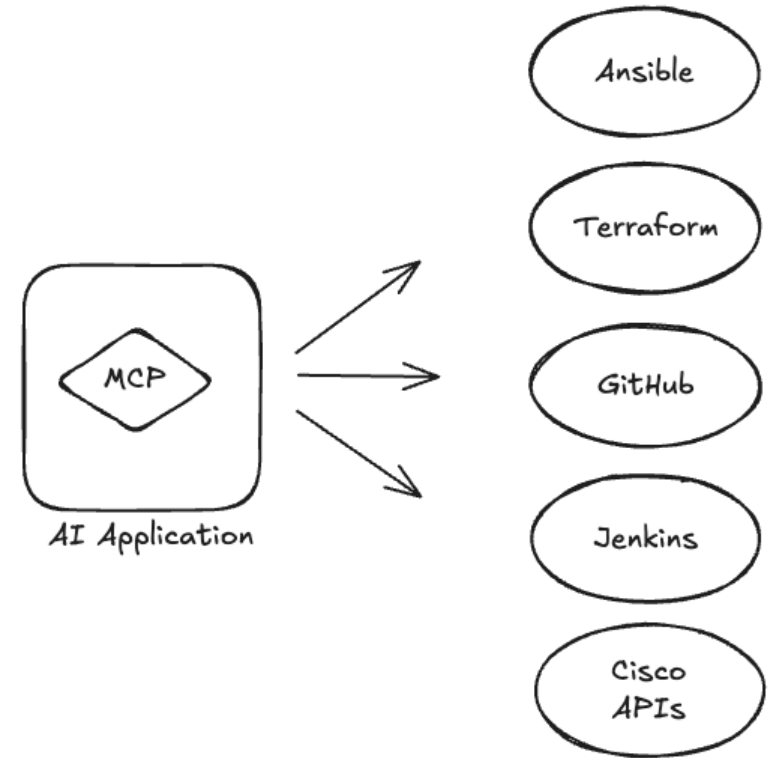
What is MCP, and Why Does it Matter?

- Model Context Protocol (MCP) is:
 - A protocol that provides a uniform way to plug in an AI model into tools and services
 - A lightweight communication protocol designed for AI agents and applications
 - Built to connect AI agents and applications to tools, APIs, databases, and filesystems
 - Structured as a client/server architecture
 - Plumbing



Operations Use Cases

- Automating Routine Tasks:
 - Generating reports, Managing GitHub repos, building Ansible playbooks, and managing CI/CD pipelines.
- Unified Data and Action Management:
 - MCP can act as a hub for interacting with diverse systems such as observability solutions from Splunk, orchestration systems such as Cisco NSO, and AI security platforms such as Cisco AI Defense.
- Enhanced Context and Decision-Making:
 - MCP-powered AI applications and agents provide richer context by accessing data from multiple sources, leading to faster, smarter decisions.
- Compliance and Security:
 - MCP interactions across your systems can be secure, compliant, and auditable when used with standardized security protocols, processes, and tools.



What is an agent?

Agents are systems that independently accomplish tasks on your behalf

Open AI

"Agent" can be defined in several ways. Some customers define agents as fully autonomous systems that operate independently over extended periods, using various tools to accomplish complex tasks. Others use the term to describe more prescriptive implementations that follow predefined workflows. At Anthropic, we categorize all these variations as agentic systems, but draw an important architectural distinction between workflows and agents:

Workflows are systems where LLMs and tools are orchestrated through predefined code paths

Agents, on the other hand, are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks

Anthropic

Realities – Agents – Are they really that special?

- Agents are code running on a machine – they consume CPU, memory, disk, network IO
- They use standard APIs
- They are automatable and programmable
- They use standard programming languages
- BUT, Agents need help:
 - Tool/Service calling – Model Context Protocol (MCP)
 - Agent-to-Agent Discovery, Composition, Deployment, Evaluation, and Communication – AGNTCY (<https://agntcy.org/>) and A2A (<https://google.github.io/A2A/>)
- Enhanced observability



© gustavofrazao / stock.adobe.com

Realities – Agents – Are they really that special?

- Agents are code running on a machine – they consume CPU, memory, disk, network IO
- They use standard APIs
- They are automatable and programmable
- They use standard programming languages
- But, Agents need help:
 - Tool/Service calling – Model Context Protocol (MCP)
 - Agent-to-Agent Discovery, Composition, Deployment, Evaluation, and Communication – AGNTCY (<https://agntcy.org/>) and A2A (<https://google.github.io/A2A/>)
 - Enhanced observability

Agents run as a process

```
ps aux | grep agent.py
shmcfarl 54290 0:08.28 /. . ./MacOS/Python agent.py
```

You can observe the process

```
ps -p 54290 -o pid,ppid,%mem,%cpu,rss,vsz
PID  PPID  %MEM  %CPU  RSS  VSZ
54290 52811  1.2   0.8   306048 412541856
```

Agents use good 'ole networking

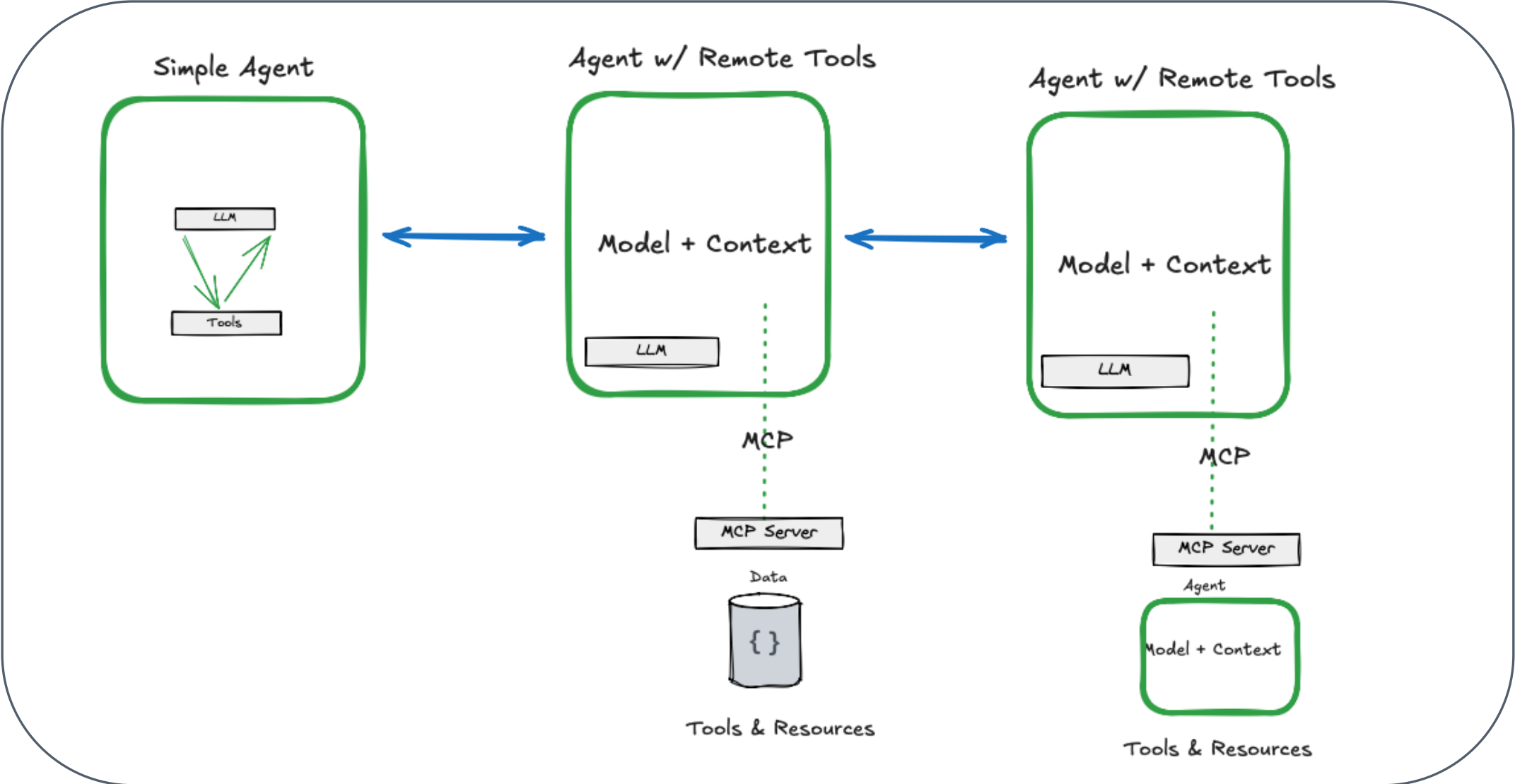
```
netstat -anv | grep 54290
tcp4      0      0  10.0.0.218.61883      209.38.174.229.4319    ESTABLISHED
tcp6      0      0  :::1.61860            :::1.1234              ESTABLISHED
```

Agent IP/Port LLM API IP/Port

Agents and LLMs use standard API interfaces

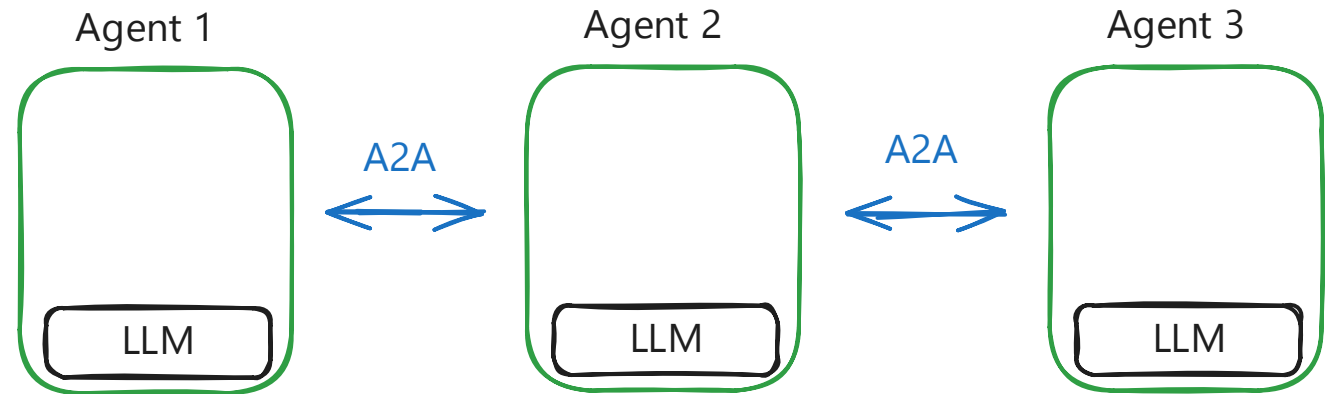
```
# curl 10.0.0.218:1234/v1/models
HTTP/1.1 200 OK
X-Powered-By: Express
Content-Type: application/json; charset=utf-8
{
  "data": [
    {
      "id": "lmstudio-community/Meta-Llama-3.1-8B-Instruct-GGUF/Meta-Llama-3.1-8B-Instruct-Q4_K_M.gguf",
      "object": "model",
      "owned_by": "organization-owner",
      "permission": [
        {}
      ]
    }
  ],
  "object": "list"
}
```

Agent to Agent Communication

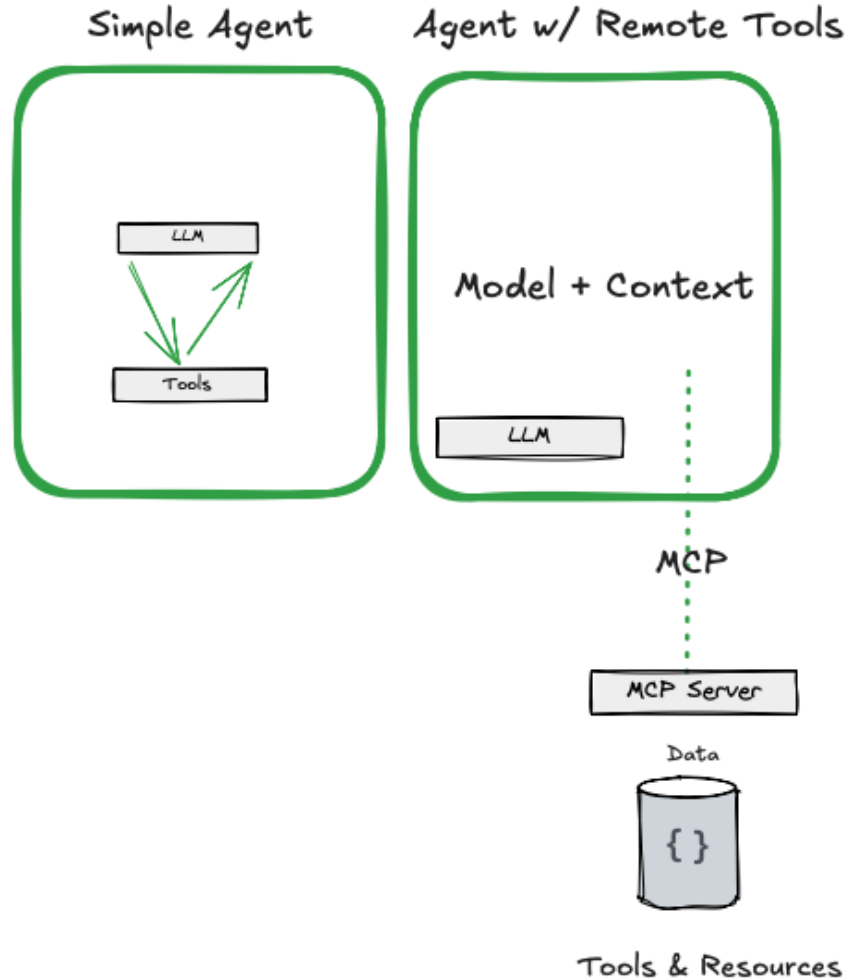


A2A: Agent2Agent

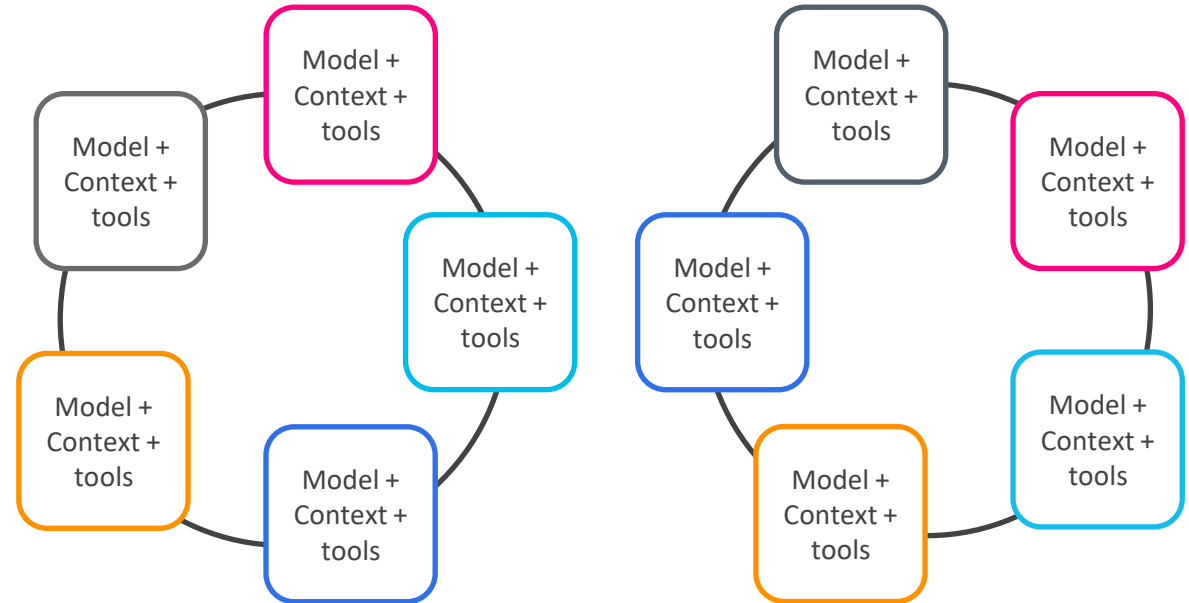
- <https://a2a-protocol.org>
- Agent2Agent Protocol: open standard developed by Google and donated the Linux Foundation
- Standardized way to enable communication between AI Agents:
 - Interoperability
 - Autonomy
 - Reduced complexity
 - Increased secure collaboration
 - Support for agentic flows (long-running operations)
- A2A is just one piece of the puzzle



Multi-agent software development



Compositions of Agents



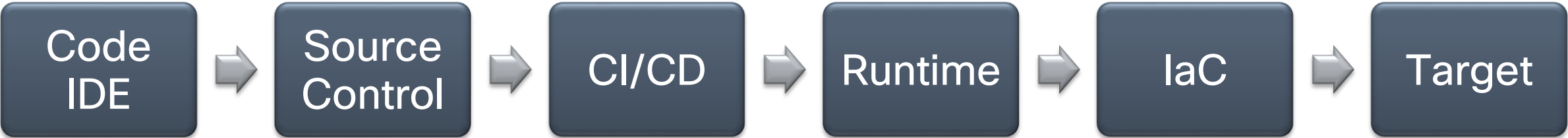
Principles:

- Agent & MCP Server reuse
- Each Agent has a specialty and is good at certain tasks
- Composition can occur by a human during MAS development or by a system at runtime!

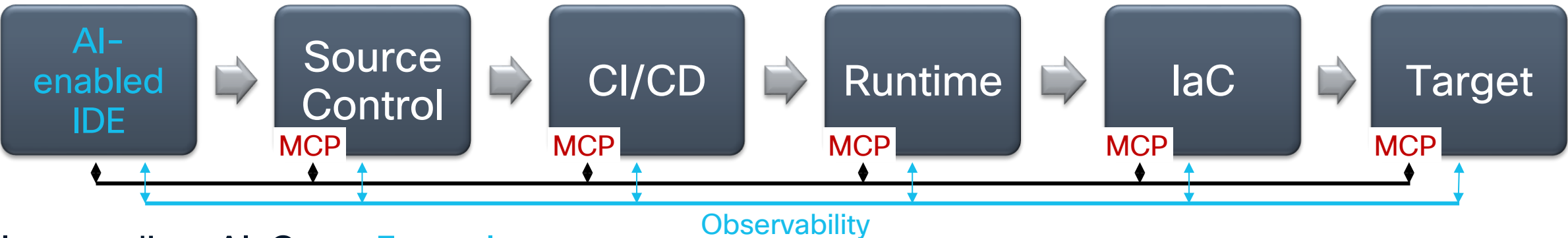
Use Cases

AI-first Ops: Example

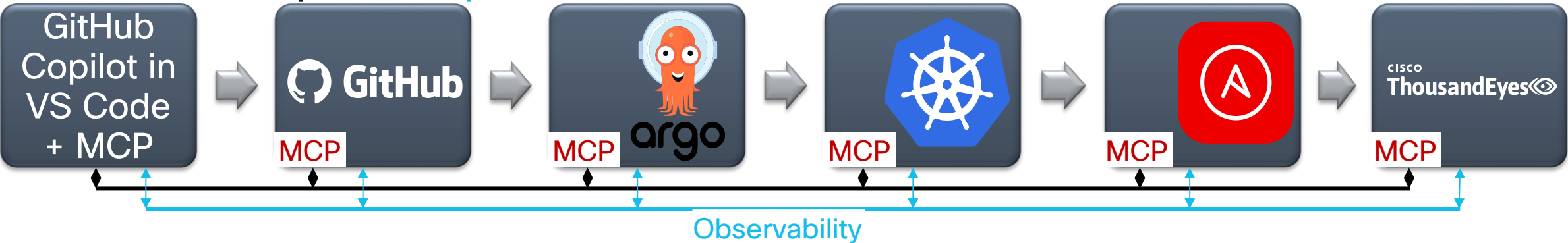
No AI-Ops



Intermediate AI-Ops

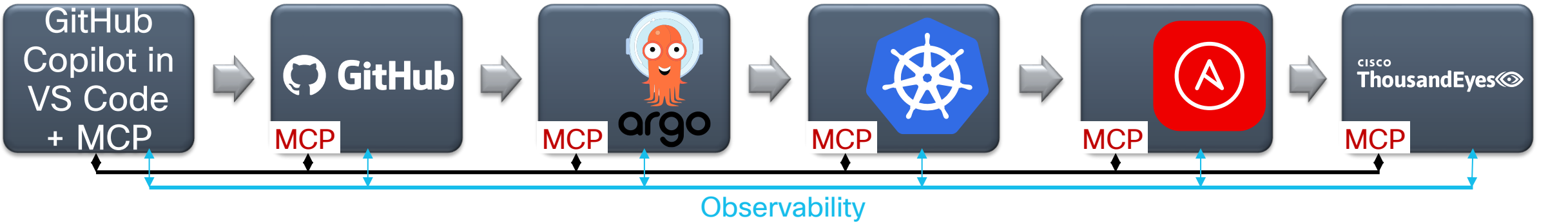
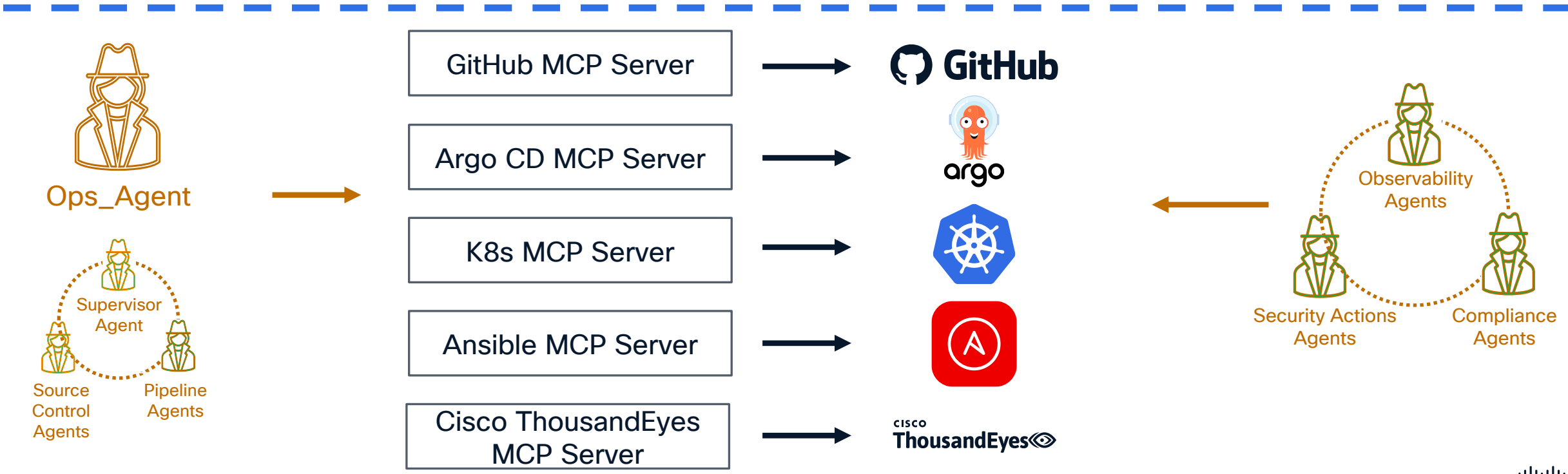


Intermediate AI-Ops - Example



AI-first Ops: Focusing on the “AI-first” part

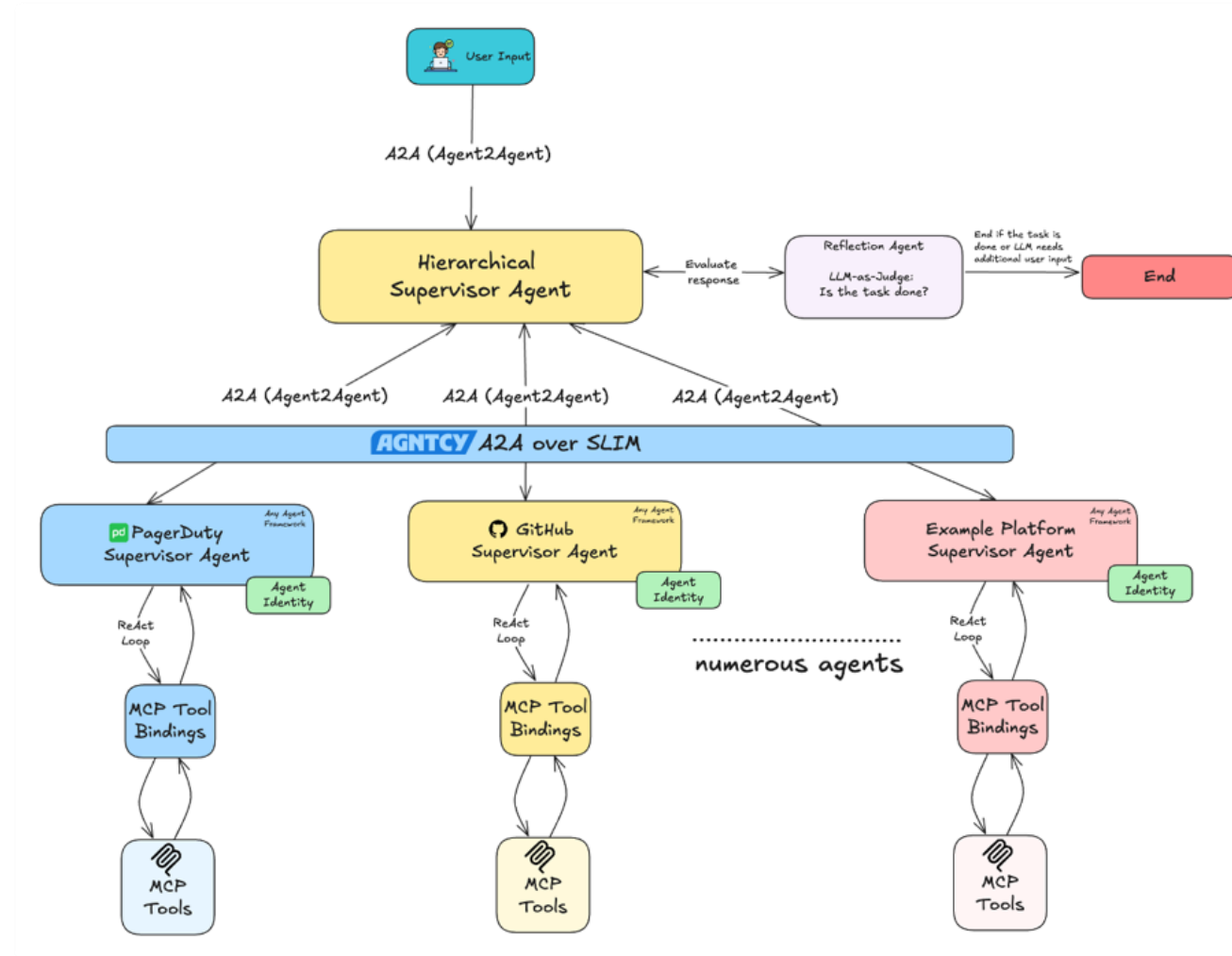
Intermediate AI-Ops - Example



Next-Gen Ops using CAIPE

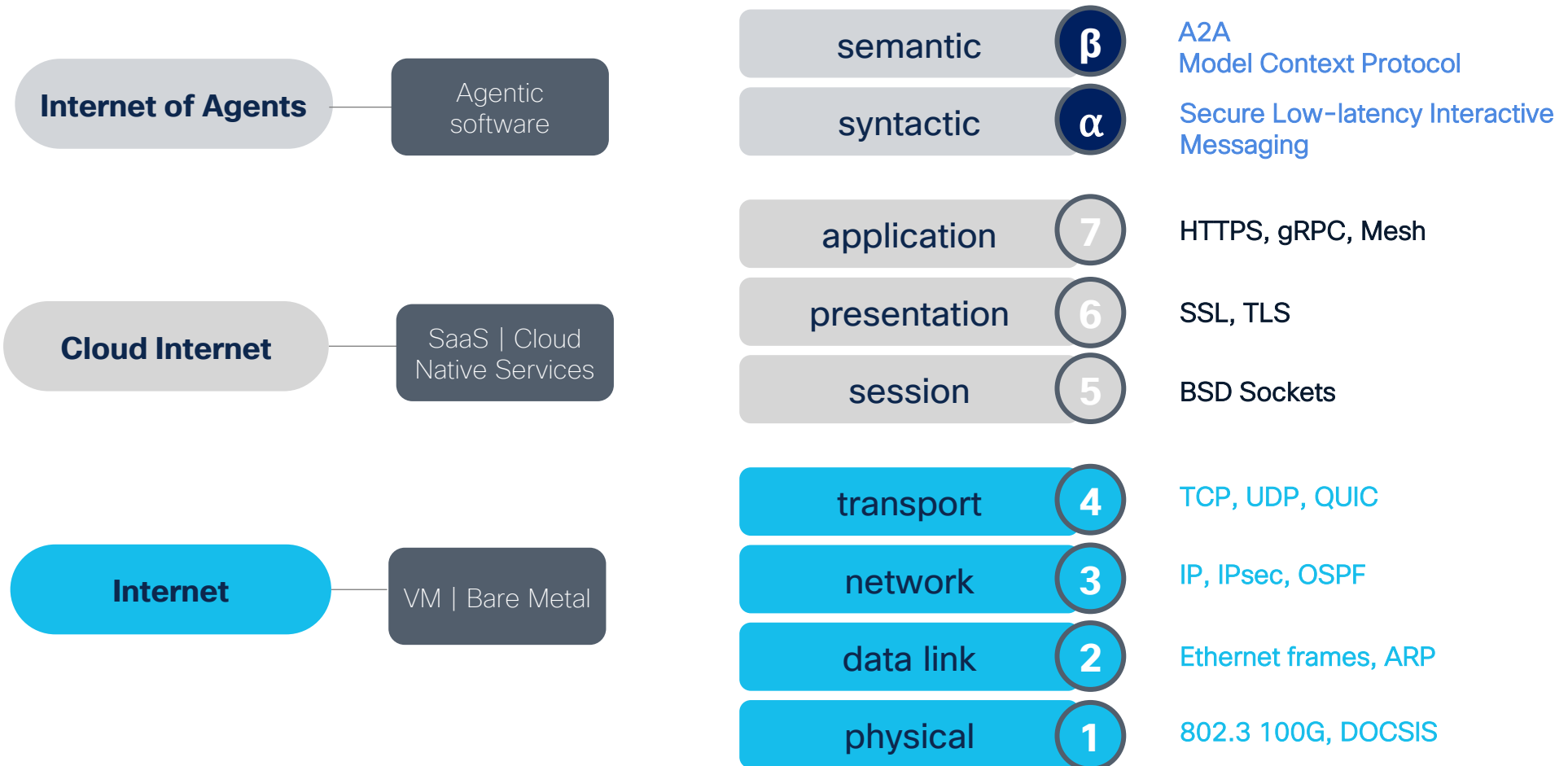
- CAIPE (Community AI Platform Engineering): “CAPE”, An open source, Multi-Agentic AI System (MAS) championed by the Cloud Native Operational Excellence (CNOE) forum.
- <https://cnoe-io.github.io/ai-platform-engineering/>
- Specialized sub-agents
- Integrate with essential engineering and operational tools:
 - ArgoCD Agent
 - PagerDuty Agent
 - GitHub Agent
 - Jira/Confluence Agent
 - Kubernetes Agent
 - Slack/Webex Agents
 - ...

Awaiting demo flow

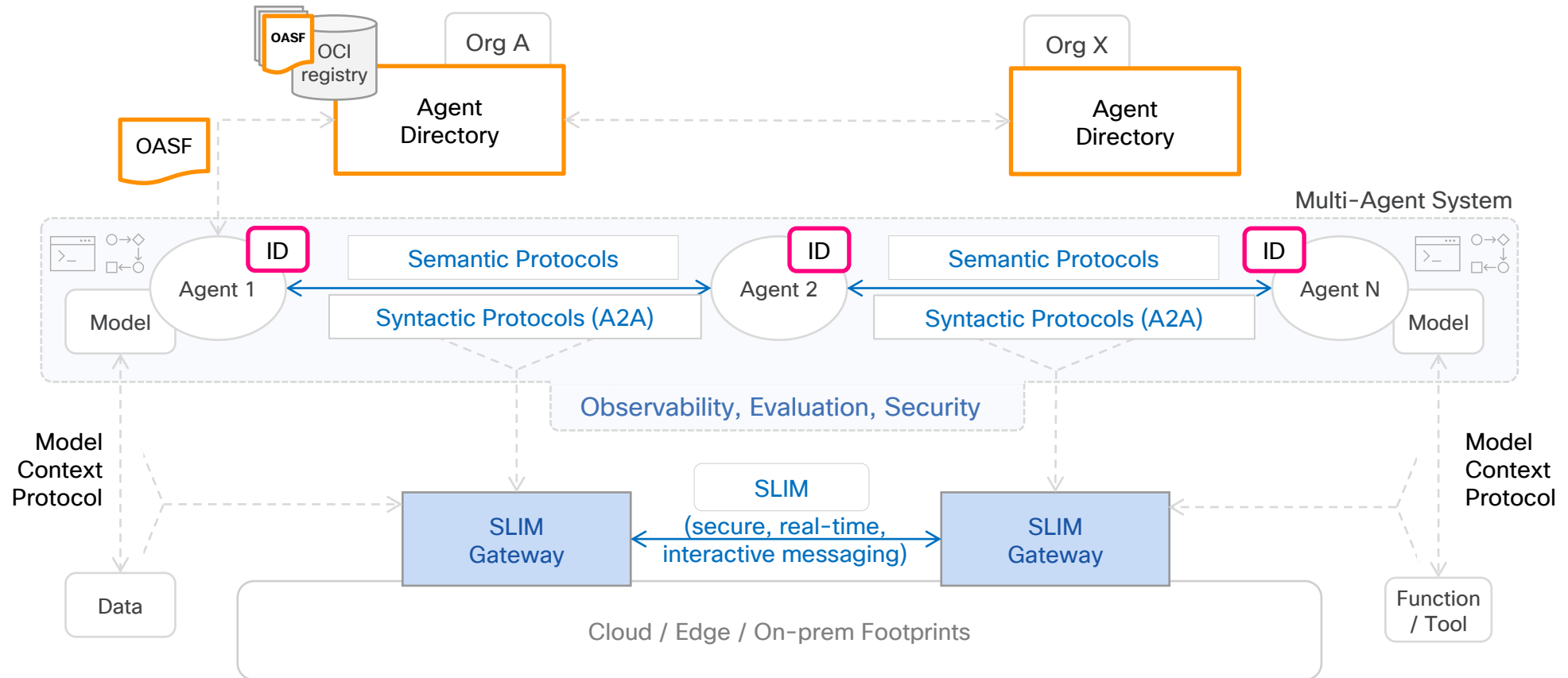


Summary

Evolving the OSI model



Internet of Agents Architecture - <https://agntcy.org/>



Summary – Call to Action

- Get Started:
 - Test the waters with Model Context Protocol: <https://modelcontextprotocol.io>
 - Reduce the FUD of agents with A2A: <https://a2a-protocol.org>
 - Simplify things with the Community AI Platform Engineering (CAIPE) project: <https://cnoe-io.github.io/ai-platform-engineering/>
 - MCP for Ops – Architecture: <https://youtu.be/Qdms0EHwhOw?si=gLWo9OR3VLk7yXA2>
 - MCP for Ops – Use cases: https://youtu.be/NXFzZsvhJR8?si=vumJRU3_9ko-IDF7
 - MCP for Ops – Cursor + MCP + Automated Pipelines: https://youtu.be/fGQ_gI_RQ_M?si=v0t8_sililPwMd7f
- Learn more about AGNTCY.org and the Internet of Agents
 - <https://agntcy.org/>
 - <https://github.com/agntcy>
- Ramp up your end-to-end game with Cisco's AI infrastructure, Security for AI, Observability for AI, and Services: <https://www.cisco.com/site/us/en/solutions/artificial-intelligence/index.html>

Complete your session evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to claim a Cisco Live T-Shirt.



Earn up to 800 points by completing all surveys and climb the Cisco Live Challenge leaderboard.



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live Events app.

Continue your education



Visit the Cisco Stand for related demos



Book your one-on-one Meet the Expert meeting



Attend the interactive education with Capture the Flag, and Walk-in Labs



Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Thank you

CISCO Live !

