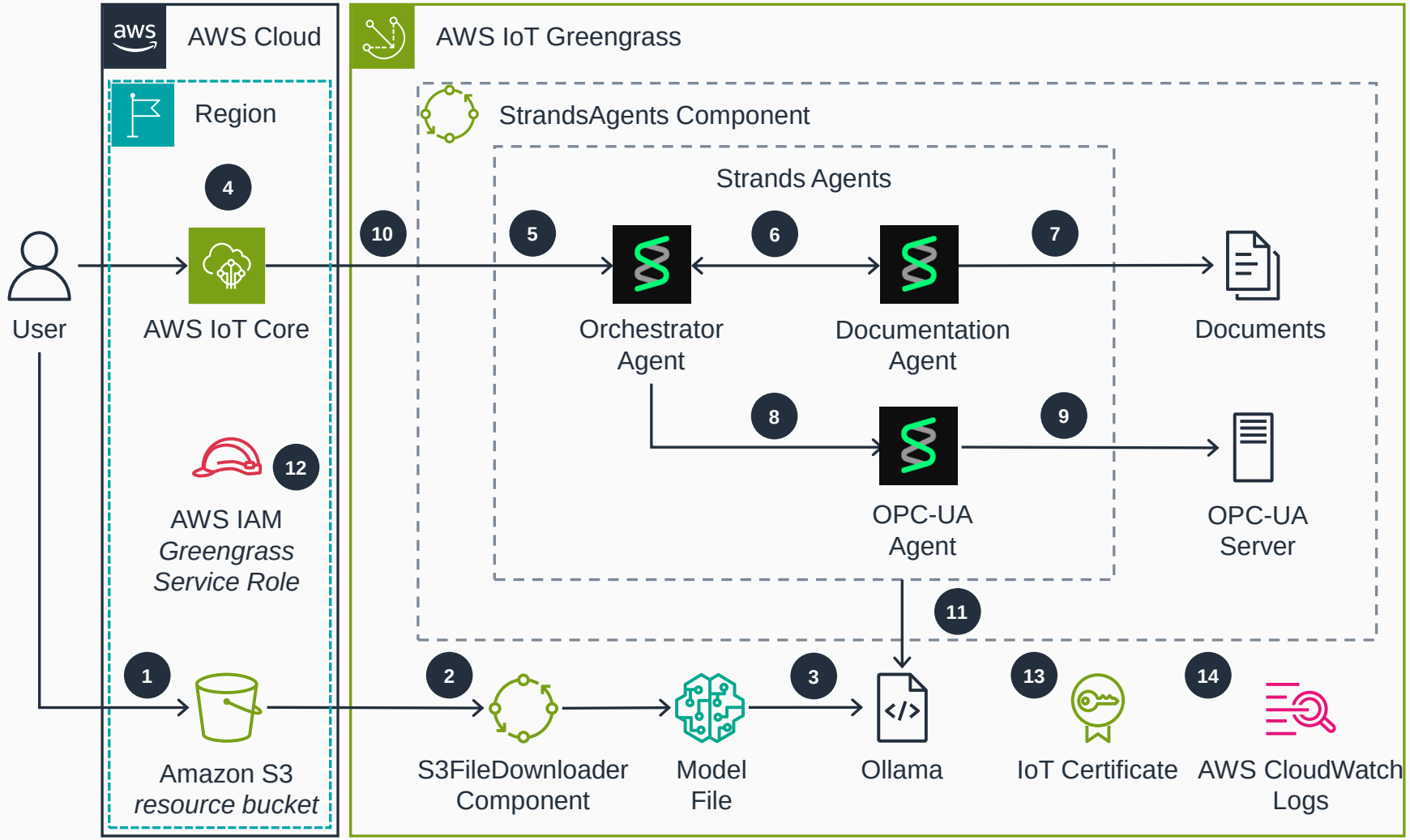
Guidance for Deploying AI Agents to Device Fleets Using AWS IoT Greengrass

This architecture diagram illustrates how to deploy AI agents to edge devices at scale. It shows the key components and their interactions, providing an overview of the architecture's structure and functionality.



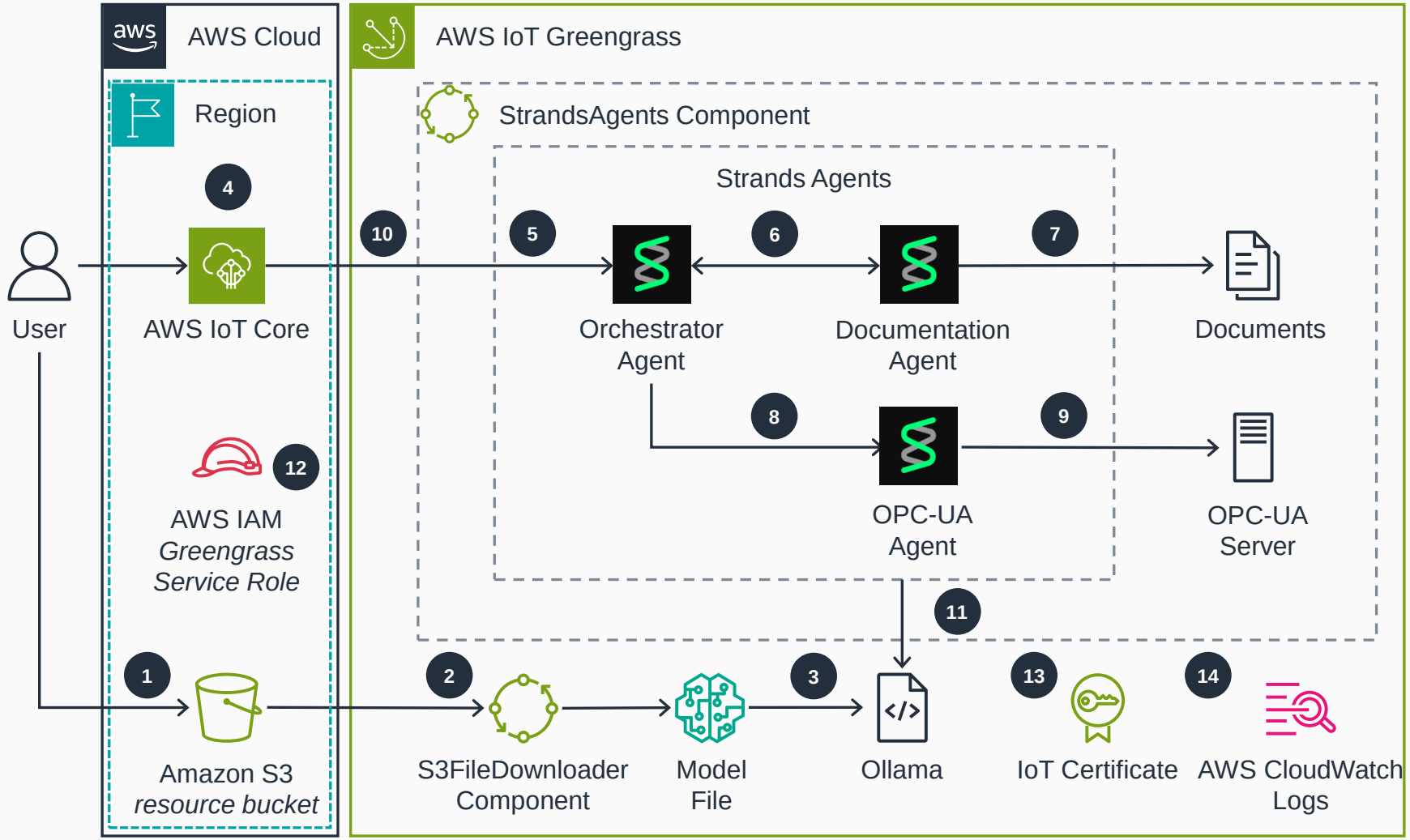
1. The user uploads a model file in GPT-Generated Unified Format (GGUF) to an **Amazon Simple Storage Service (Amazon S3)** bucket which **AWS IoT Greengrass** devices have access for.
2. The devices in the fleet receive a file download job. The S3FileDownloader component processes this job and downloads the model file to the device from the **Amazon S3** bucket.
3. The model file in GGUF format loads into Ollama when the StrandsAgents component makes the first call to Ollama. The model name is specified in the recipe.yaml file of the component.
4. The user sends a query to the local agent by publishing a payload to a device-specific agent topic in **AWS IoT Core** MQTT broker.
5. After receiving the query, the component leverages the Strands Agents SDK's model-agnostic orchestration capabilities. The Orchestrator Agent perceives the query, reasons about the required information sources, and acts by calling the appropriate specialized agents (Documentation Agent, OPC-UA Agent, or both) to gather comprehensive data before formulating a response.
6. If the query is related to information that can be found in the documentation, Orchestrator Agent calls Documentation Agent.
7. Documentation Agent finds the information from the provided documents and returns it to Orchestrator Agent.
8. If the query is related to current or historical machine data, Orchestrator Agent will call OPC-UA Agent.
9. OPC-UA Agent makes a query to the OPC-UA server depending on the user query and returns the data from server to Orchestrator Agent.

Reviewed for technical accuracy October 14, 2025

AWS Reference Architecture

Guidance for Deploying AI Agents to Device Fleets Using AWS IoT Greengrass

This architecture diagram illustrates how to deploy AI agents to edge devices at scale. It shows the key components and their interactions, providing an overview of the architecture's structure and functionality.



10 Orchestrator Agent forms a response based on the collected information. The StrandsAgents component publishes the response to a device-specific agent response topic in the **AWS IoT Core** MQTT broker.

11 The Strands Agents SDK enables the system to work with locally deployed foundation models through Ollama at the edge, while maintaining the option to switch to cloud-based models like those in **Amazon Bedrock** when connectivity is available.

12 The **AWS Identity and Access Management (IAM)** Greengrass Service Role provides access to the **Amazon S3** resource bucket to download models to the device.

13 The IoT certificate attached to the IoT thing allows the StrandsAgents component to receive and publish MQTT payloads to **AWS IoT Core**.

14 The **IoT Greengrass** component logs the component operation to the local file system. Optionally, **AWS CloudWatch** Logs can be enabled to monitor the component operation in the **CloudWatch** console.

Reviewed for technical accuracy October 14, 2025

AWS Reference Architecture