WP3 Node Architecture & Building Blocks

WP3T34-03: Versal Model deployment layer

Developed by: PLC2

Component description

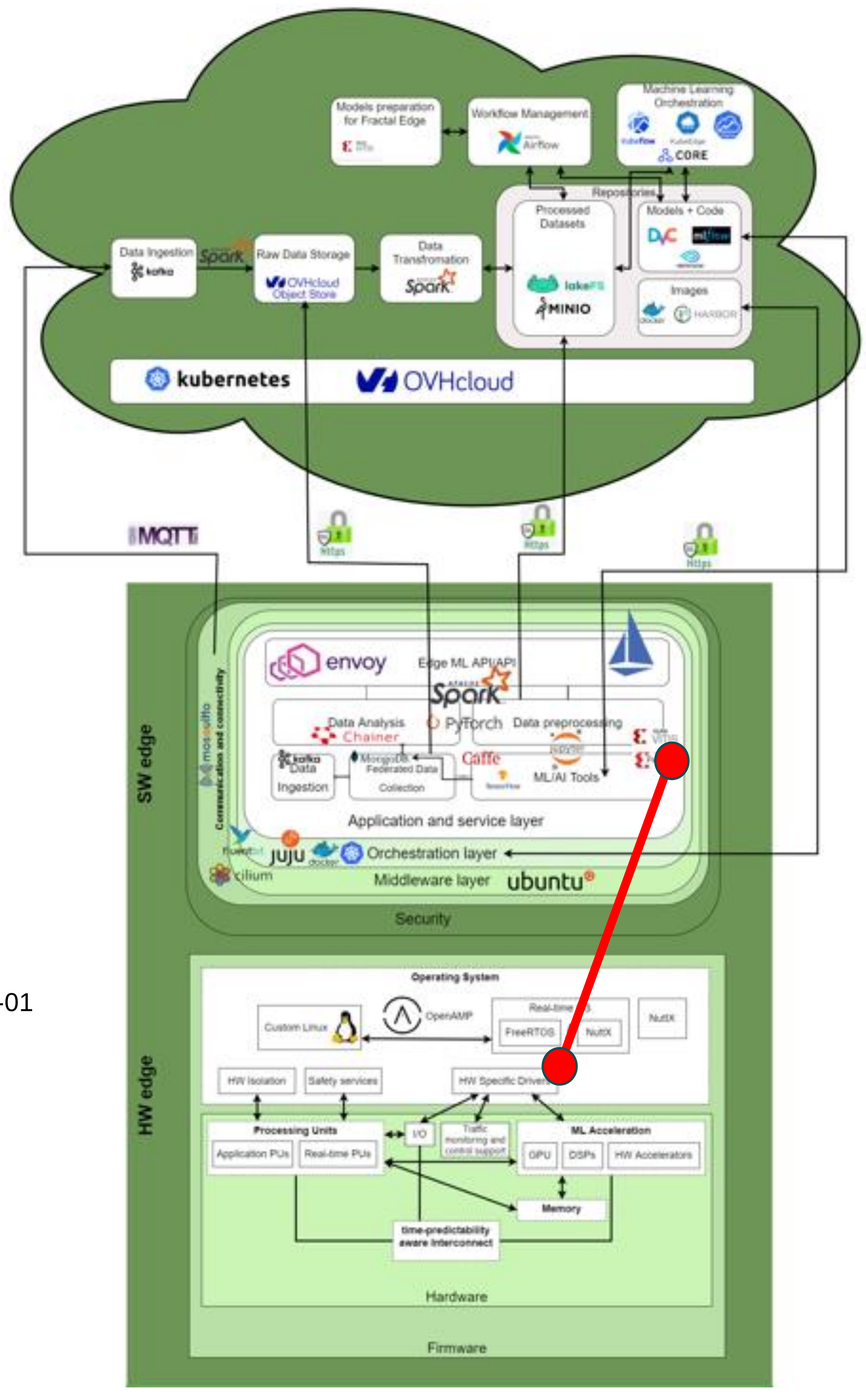
- Objective of the component:
  Model deployment on the Versal APU + DPU control from model repository images

- Fractal Features associated:
  ADAPTABILITY [AI]
  FRACTALITY [MODEL]

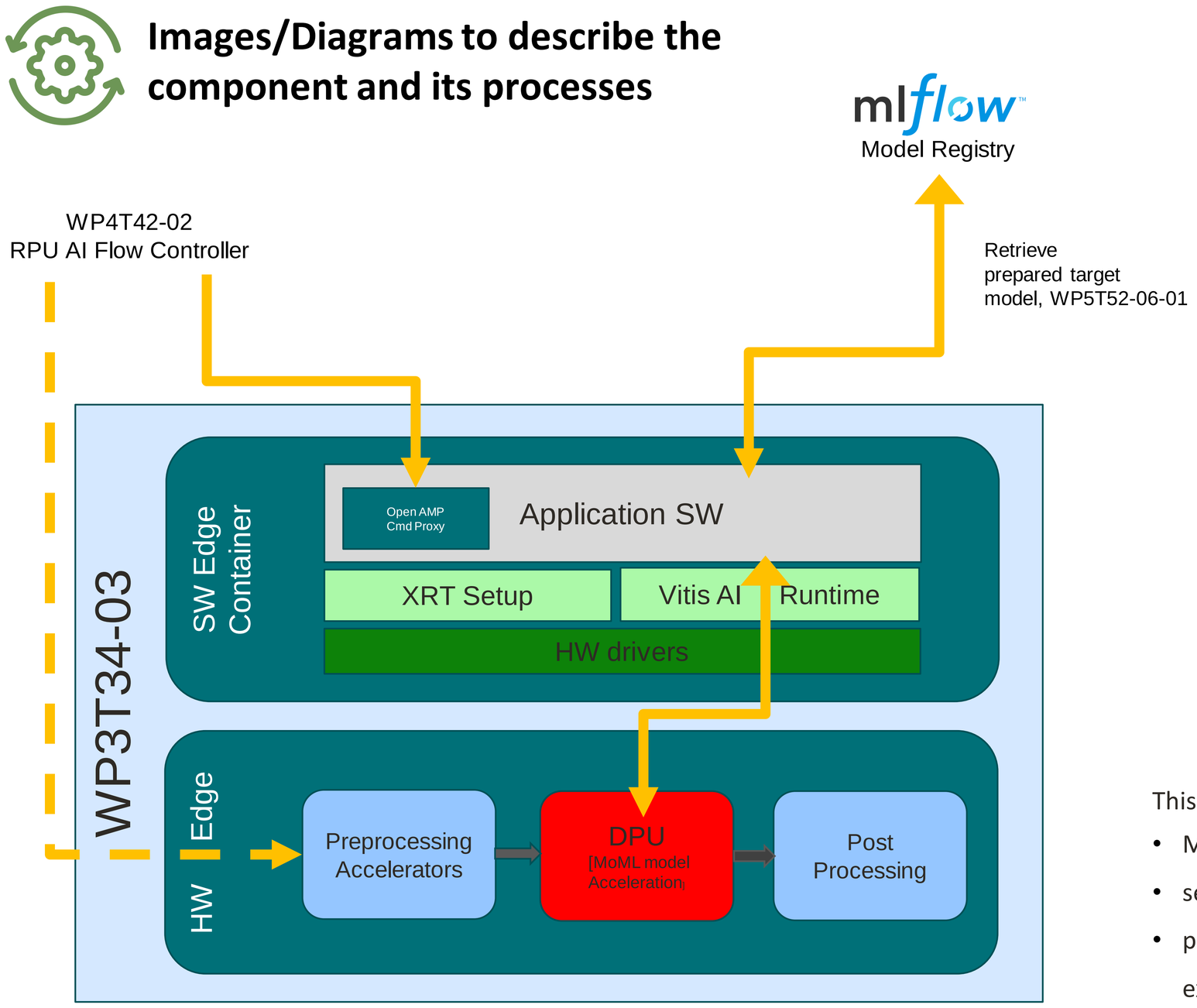
- Inputs/Outputs:
  Data Source as defined by the AI model (HW path or collection)
  Decision derived from local inference

- Integration: WP4T41-02 (not limited to)

Component location



Images/Diagrams to describe the component and its processes



This component provides a containerized service that serves

- ML model target code retrieval from MLflow model registry
- setup and run the target specific ML accelerator
- provide proxy services into the active accelerator to handle external flow control

Get started

Retrieve the component's container image into the service enhanced FRACTAL platform in Versal VCK190 by providing it in the boot image or downloading it from an image regsitry through orchestration.

Start containerized service within the local cluster. Current state will wait for external connection.

A connection client (see WP4T41-02) provides protrocol based actions:

- Loading and discarding a ML model
- Setting up the ML model
- Executing ML inference (repetitive, single, blocking, non-blocking)
- Sendig out inference estimates

EU2020 Horizon

Project N.877056

ECSEL Joint Undertaking
Electronic Components and Systems for European Leadership

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 877056. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Italy, Austria, Germany, Finland, Switzerland.