

This work has been funded by the Horizon Europe research and innovation programme of the European Union, under grant agreement no 101092912, project MLSysOps.

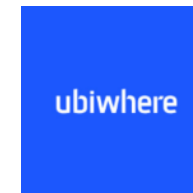
MLSysOps (introduction)

Spyros Lalis

University of Thessaly



NTT DATA



Overview

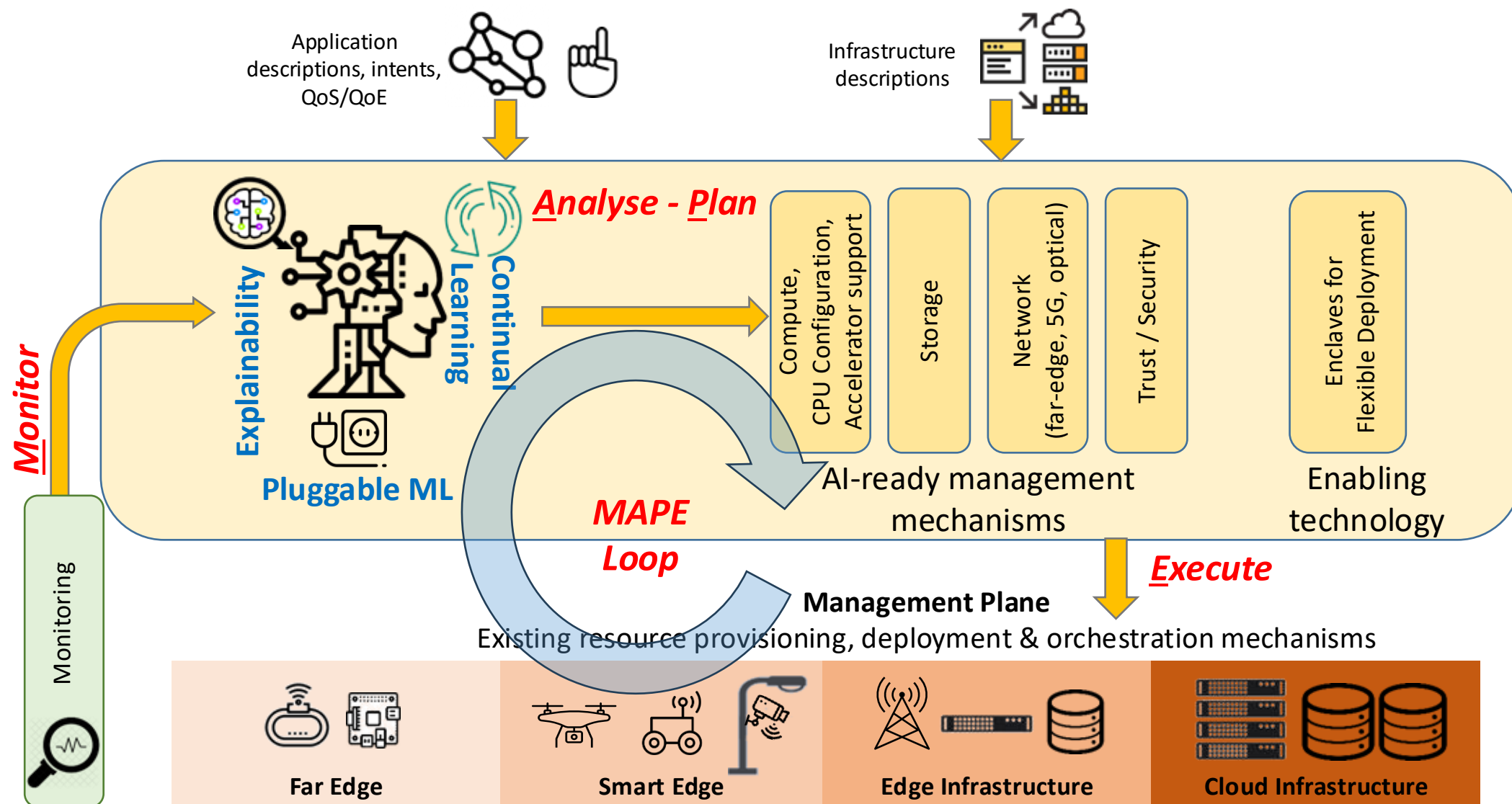
Goals

- **Autonomic** system management and configuration in the **cloud-edge-IoT** continuum
- Explore different aspects: **deployment**, **computing/acceleration**, **storage**, **communication**, **trust**
- Using **AI/ML methods**, while disassociating **mechanisms** from **control decisions**

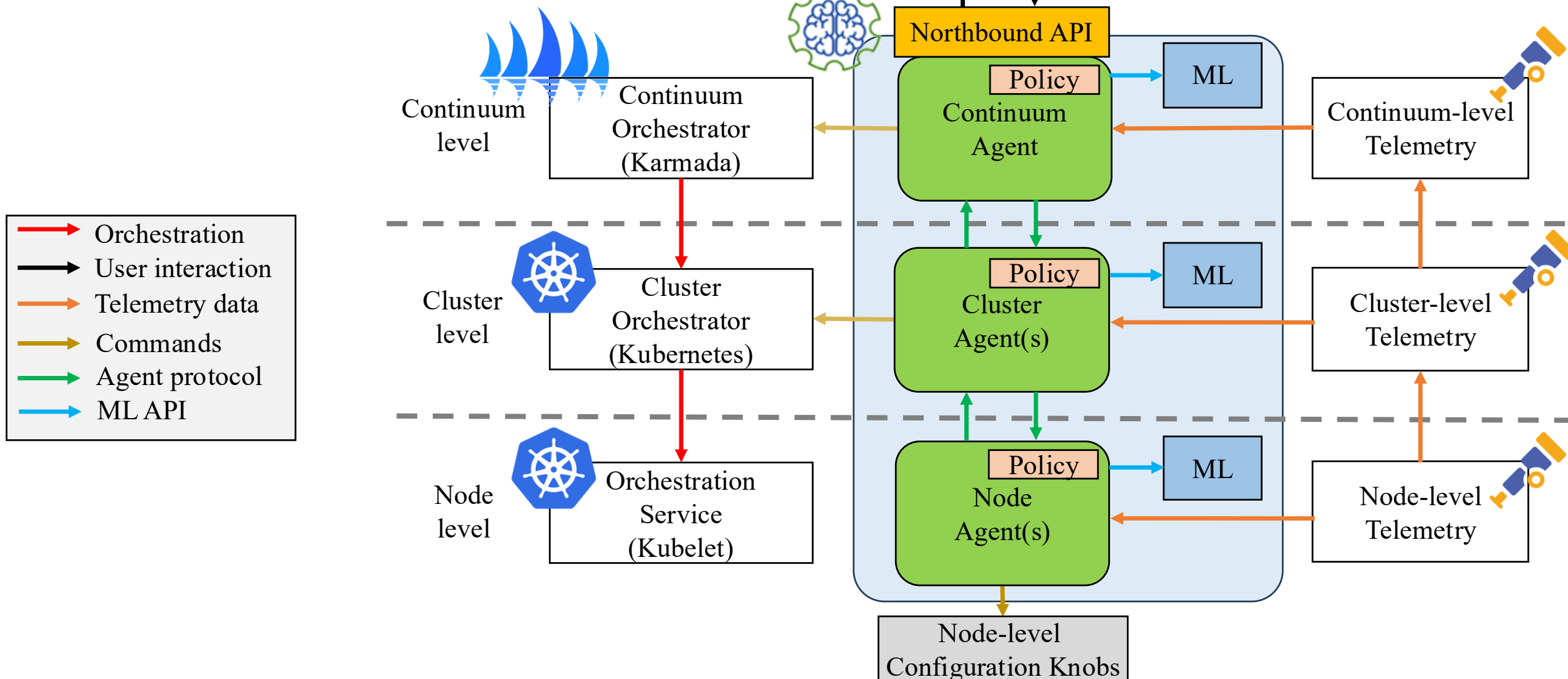
Main features of the MLSysOps framework

- Support for **distributed** applications consisting of **multiple (interacting) components**
- Open **hierarchical agent-based framework** for scalable and adaptive control
- Flexible AI architecture supporting both **plug-in policies** & **ML models**
- Ability to produce/consume both **system-** and **application-level performance metrics**
- Standard deployment and telemetry technologies

High-level concept

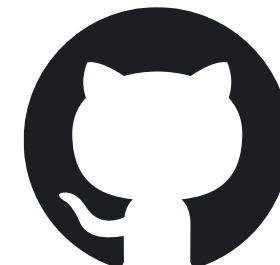


Framework Architecture



Open-source release

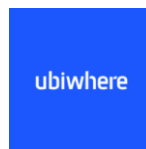
- Core MLSysOps framework released as free open-source software
github.com/mlsysops-eu/mlsysops-framework
 - **Kubernetes-native deployment** using CRDs
 - **Multi-cluster orchestration** powered by Karmada
 - **Dynamic telemetry and observability layer** based on OpenTelemetry, easily extendable
 - **Plugin systems** to support configuration policies and custom mechanisms
 - **Northbound REST API** service for external applications or CLI use
 - **ML Connector Service** for model deployment, retraining, and explainability integration
 - **Support for far-edge devices and multiple container runtimes**
 - **Built-in storage and resource management** across the continuum
- Online tutorial
<https://www.youtube.com/watch?v=IGqd4HQfVRY>



Following presentations

Real application use cases

- Smart cities
- Precision agriculture



Selected functional aspects

- AI/ML in the continuum
- Storage management
- Optimized FPGA-based ML
- Acceleration at the edge

